CDF & D0 Computing

Version 0.22

4 / 9 / 2008

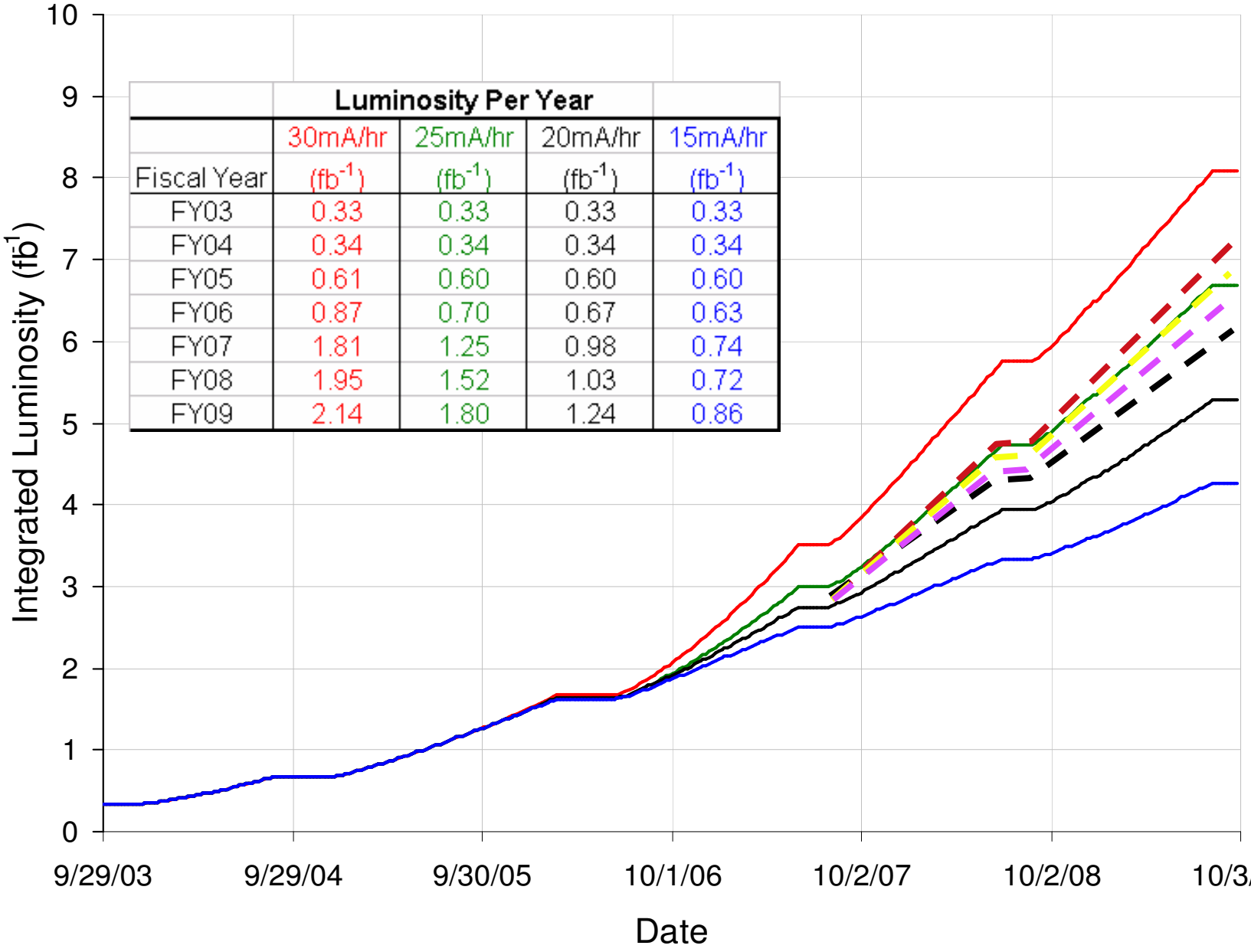
GP Yeh

Future Challenges

- Higher instantaneous luminosity
  - Larger events, slower reconstruction, tracking more difficult, need more CPU per event
- Higher integrated luminosity and higher data taking rate
  - Larger data samples
    - Need more processing power
    - Need more storage
- Migration of physicists to LHC experiments
  - Human resources for operations are shrinking
- FY2010 Running has been proposed

Luminosity Projections with Updated Model Scenarios

	Luminosity Per Year			
	30mA/hr	25mA/hr	20mA/hr	15mA/hr
Fiscal Year	(fb^{-1})	(fb^{-1})	(fb^{-1})	(fb^{-1})
FY03	0.33	0.33	0.33	0.33
FY04	0.34	0.34	0.34	0.34
FY05	0.61	0.60	0.60	0.60
FY06	0.87	0.70	0.67	0.63
FY07	1.81	1.25	0.98	0.74
FY08	1.95	1.52	1.03	0.72
FY09	2.14	1.80	1.24	0.86

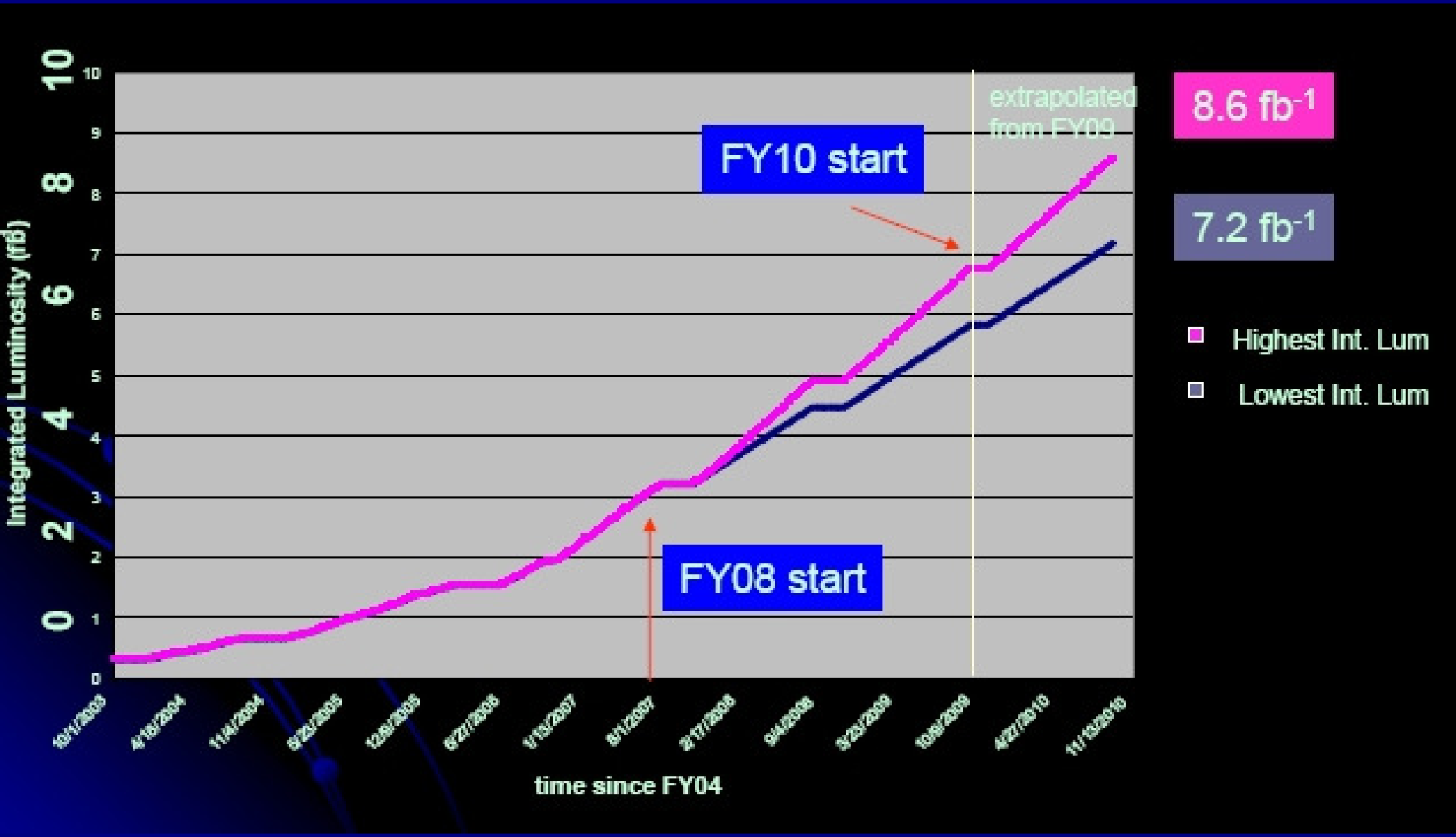


goal

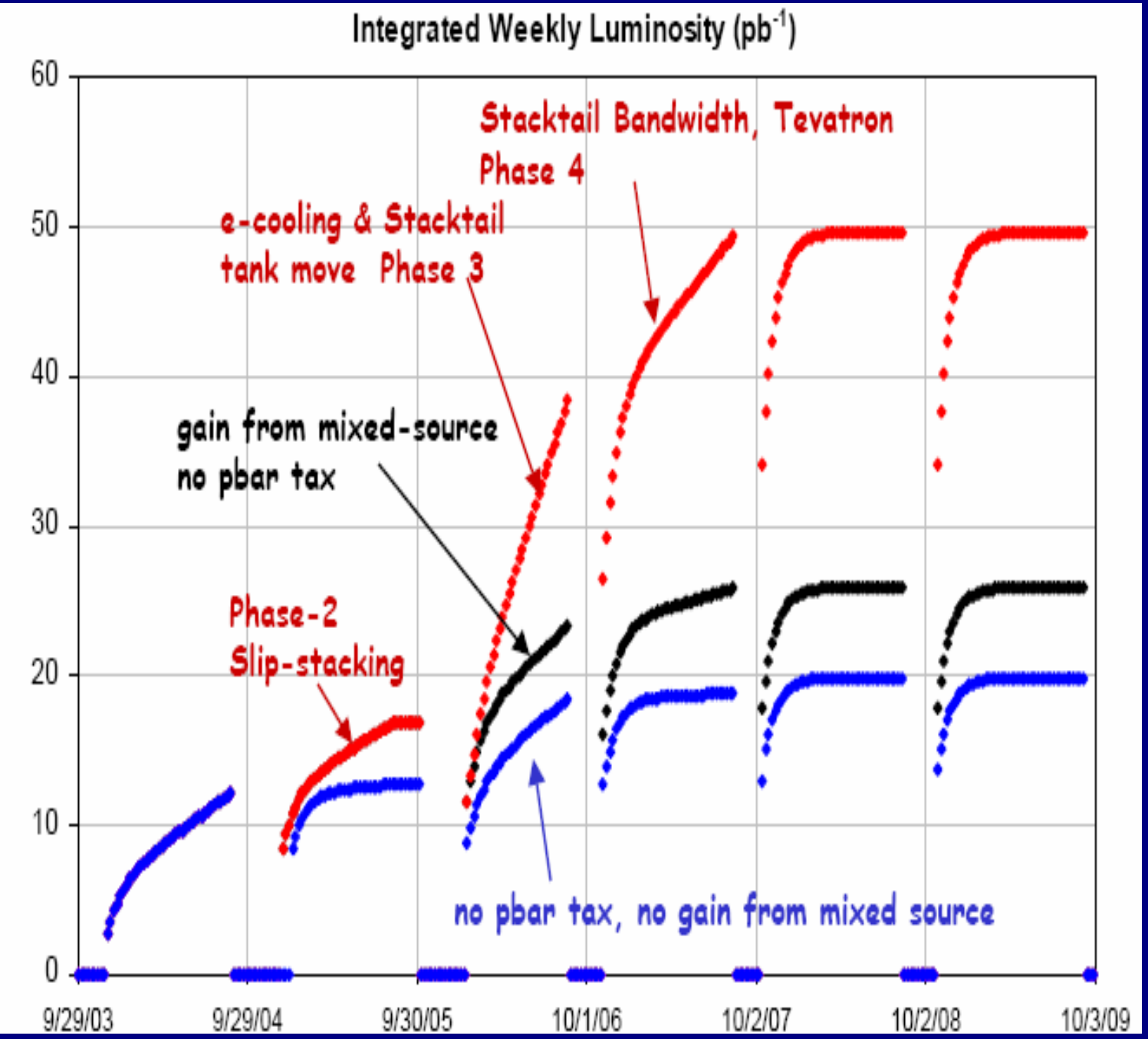
updated model scenarios 1 to 4

≈1 fb^{-1} spread in total

Integrated Luminosity



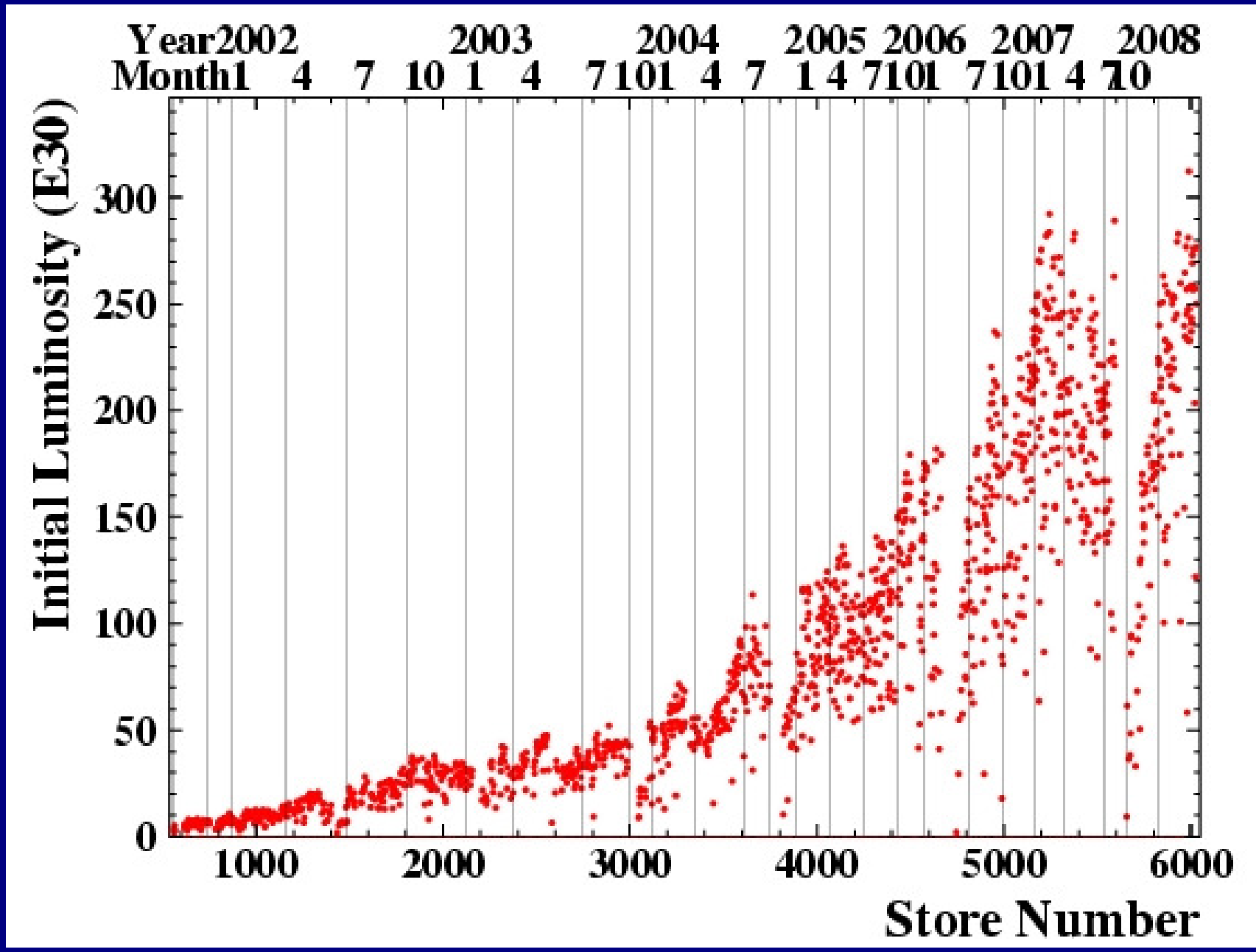
Tevatron Future Operation



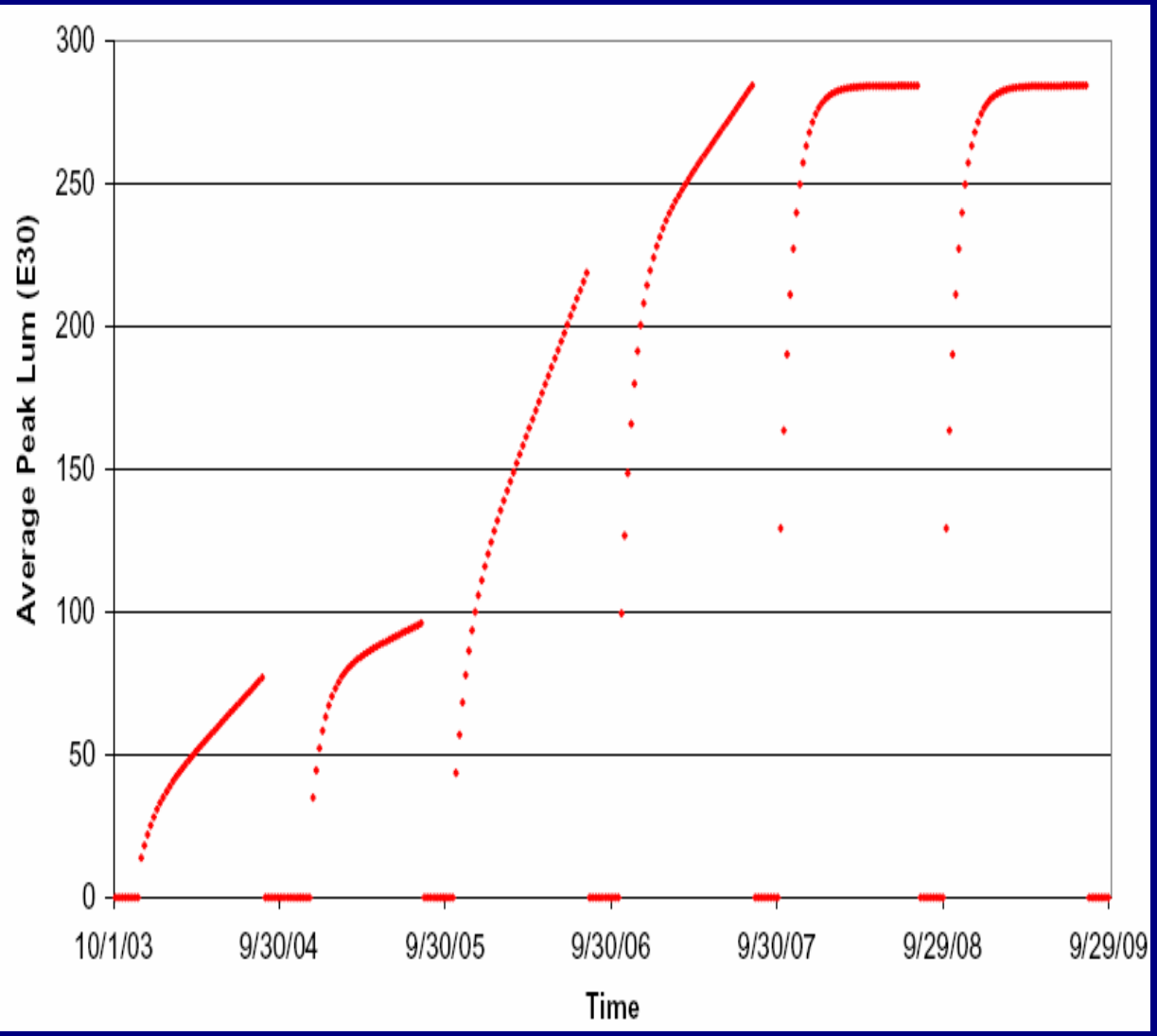
+ FY 2010 ?

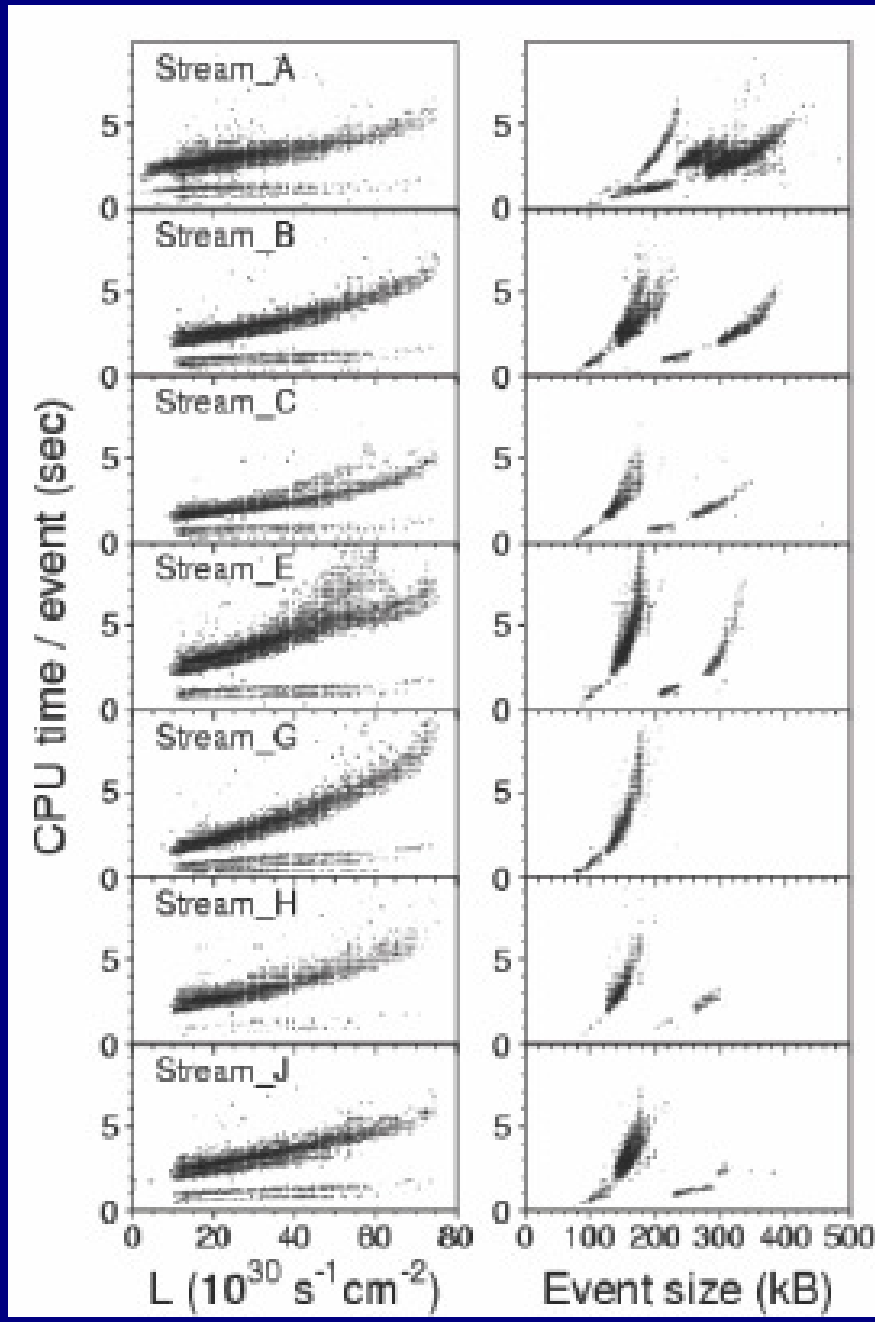
Still large factors to be gained over the next few years

Initial Luminosity



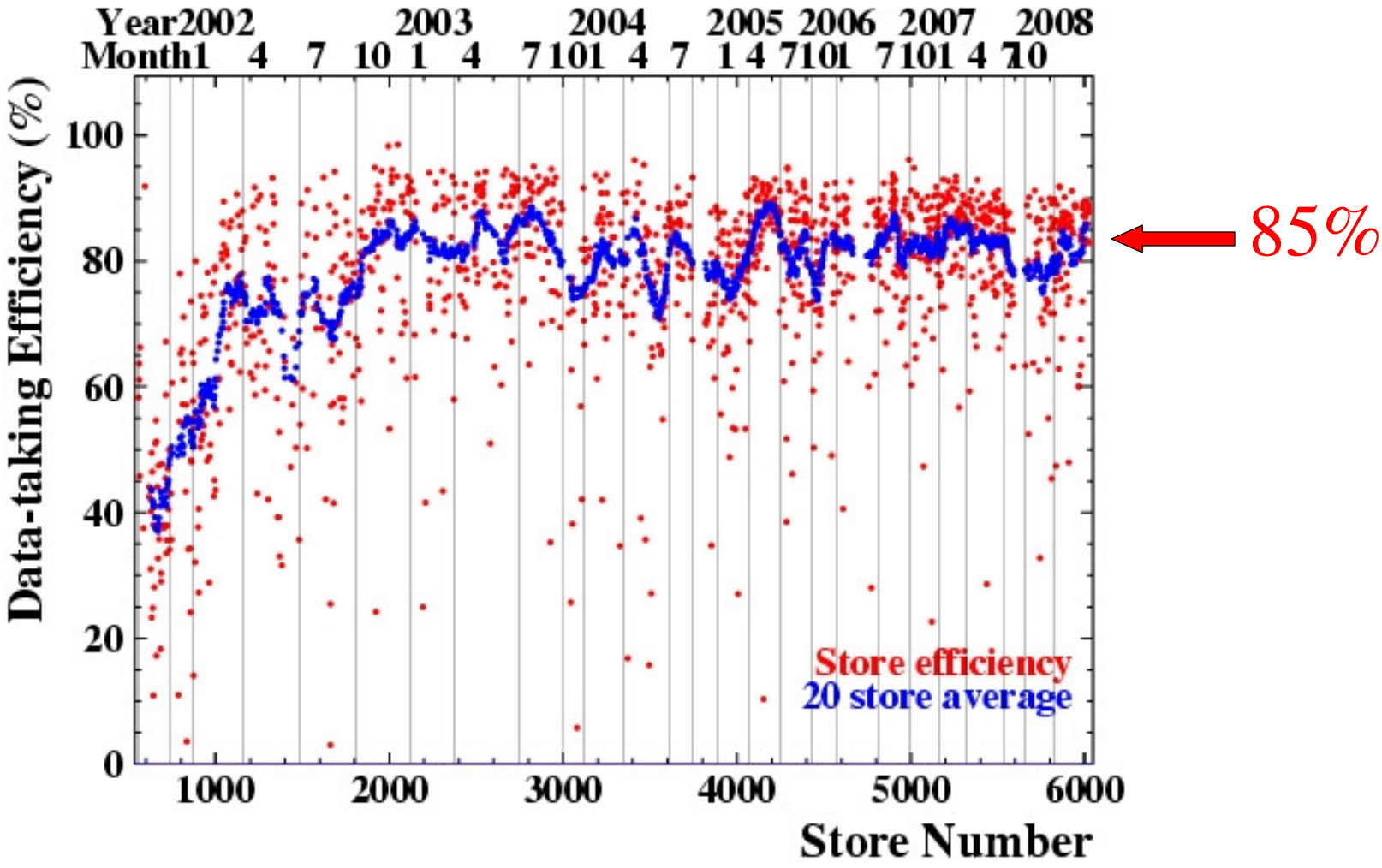
Challenge I: Higher Inst.Luminosities



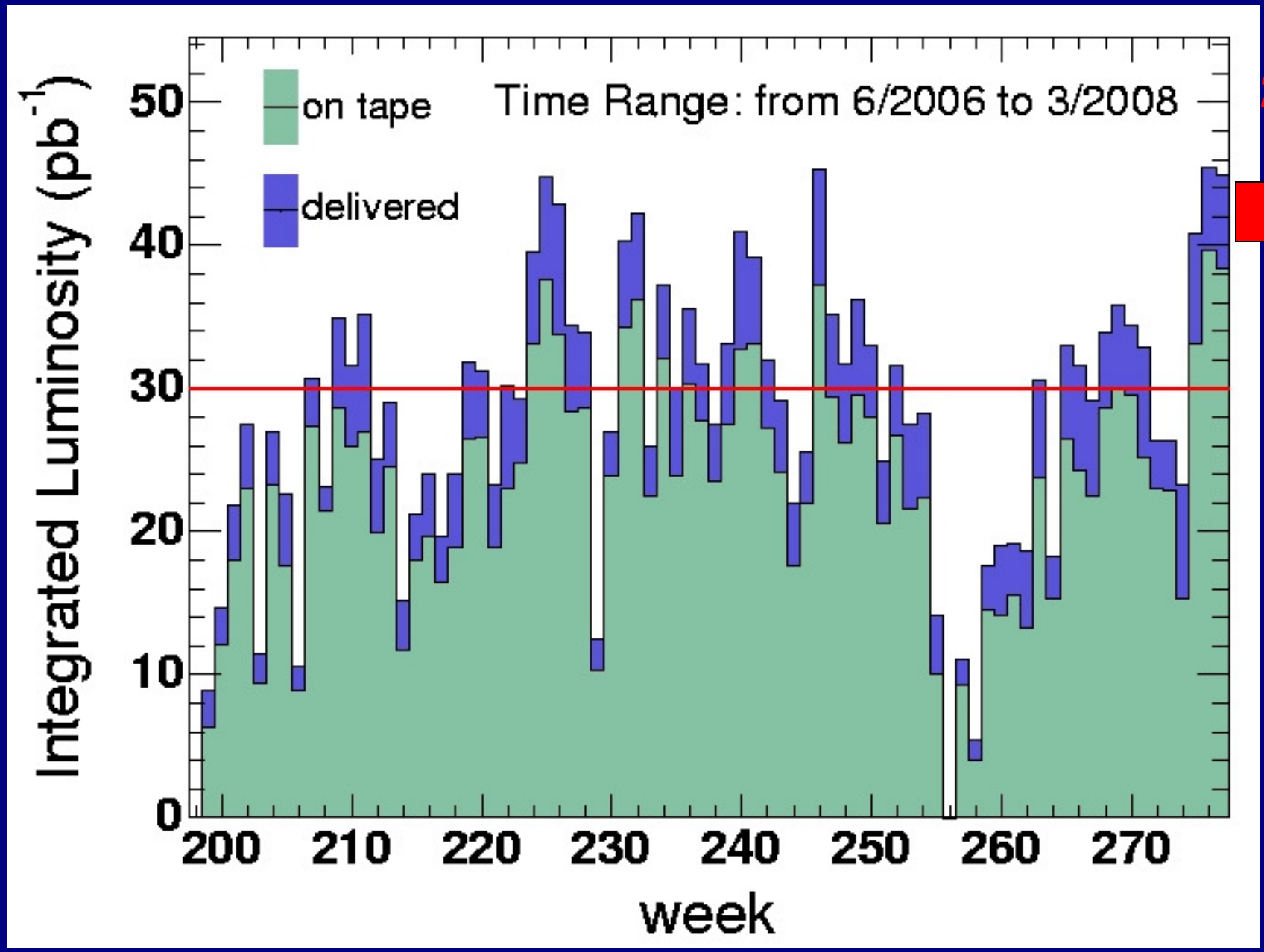


CDF

	2003		2007
Level 1 trigger:	12 KHz	⇒	35 KHz
Level 2 trigger:	300 Hz	⇒	800 Hz
Level 3 trigger:	24 MB/s	⇒	100 MB/s

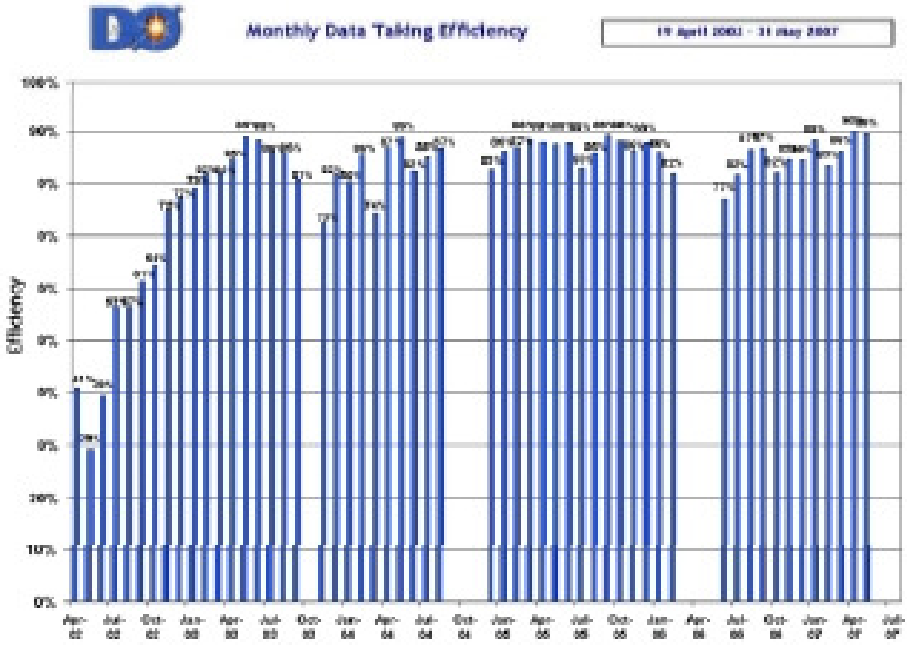


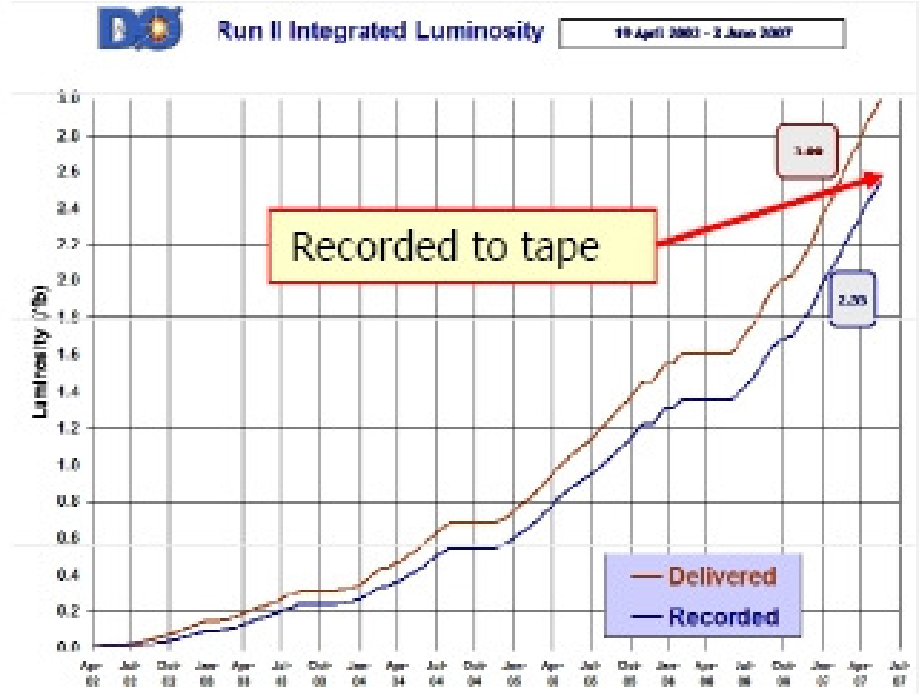
CDF weekly Integrated Luminosity



2008-2010

Operations





- The experiment is operating well and recording physics quality data
  - Typical "good" day 7 pb^{-1}
  - Typical "good" week 40 pb^{-1}
    - Run I top quark discovery in a week!
- On average 85% data taking efficiency
  - 5% are trigger/readout system disables
  - 10% are begin/end stores, failures
- As of today DØ has ~2.5 fb^{-1} on tapes
  - All detectors functioning well
  - Already reported physics results from early 2007 data

All of the above is due to the dedicated team of experts and shifters

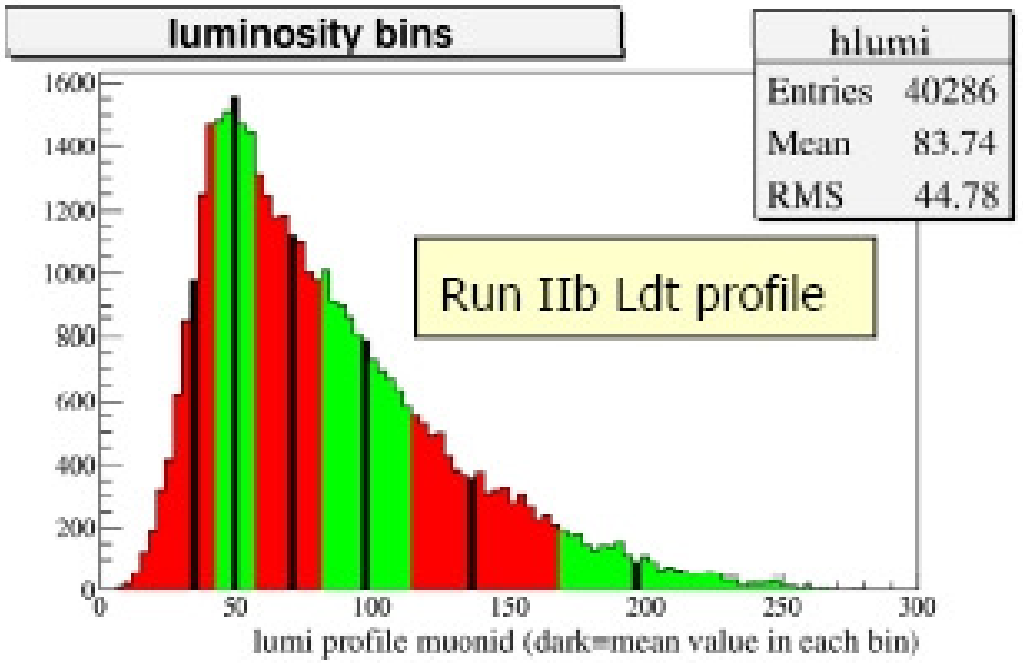
This week celebrating delivery of 3.0 fb^{-1} to the DØ experiment
Thanks to the Accelerator Division!

D0

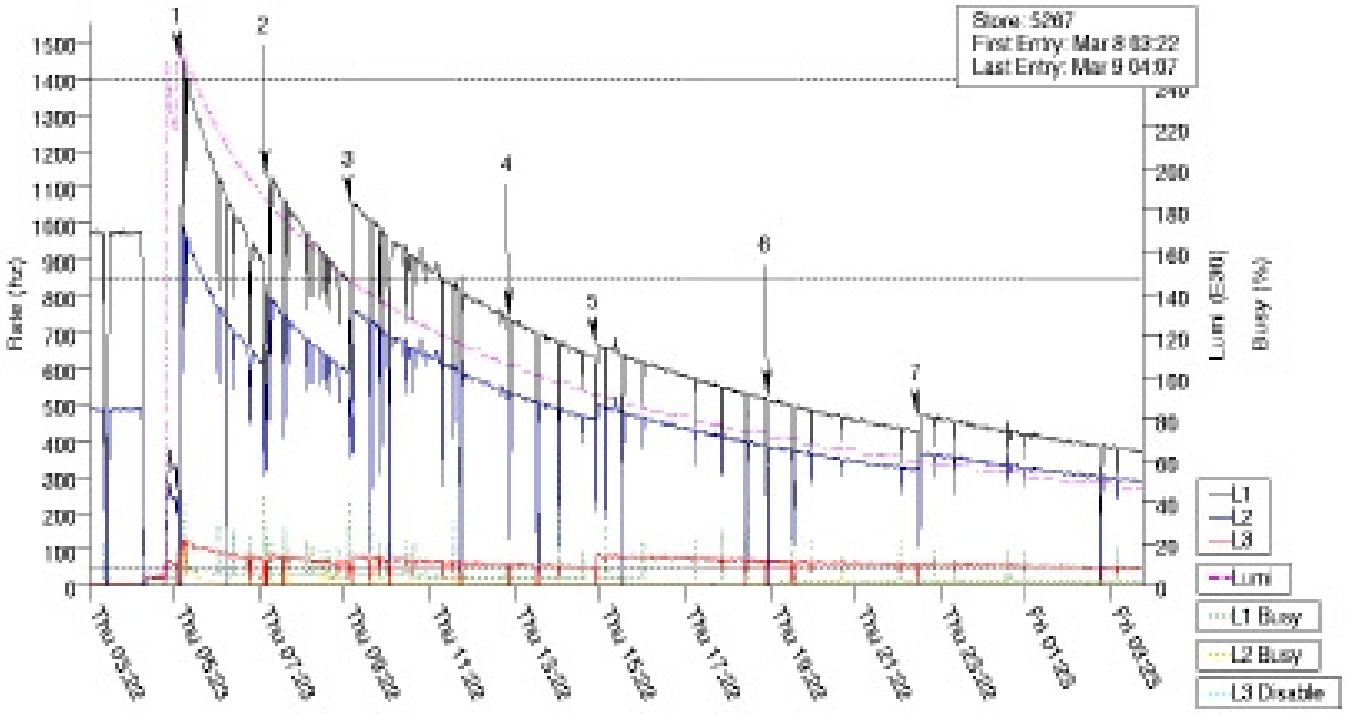
Trigger system selects ~100Hz of events
to write to tape out of ~2 MHz interactions rate
~ 10^5 rejection!

Current DØ capabilities

- Level 1 trigger ~2kHz
- Level 2 trigger ~1kHz
- Level 3 trigger ~100Hz

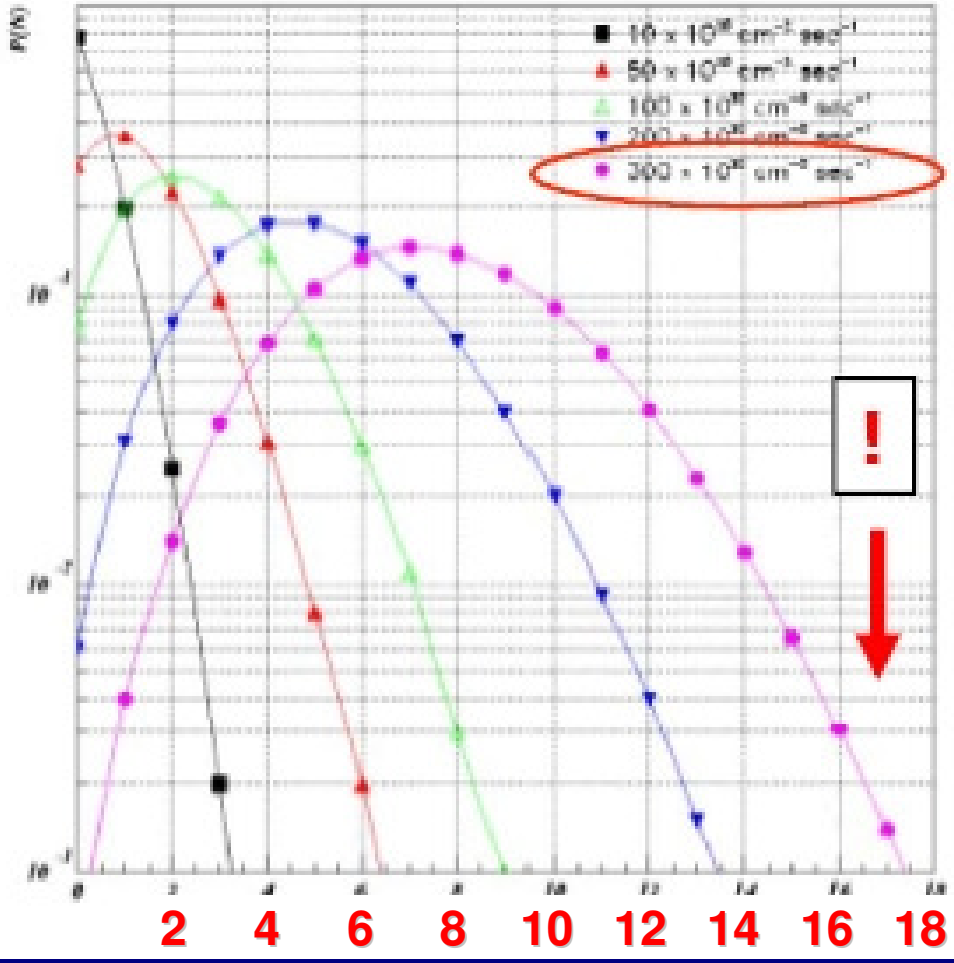


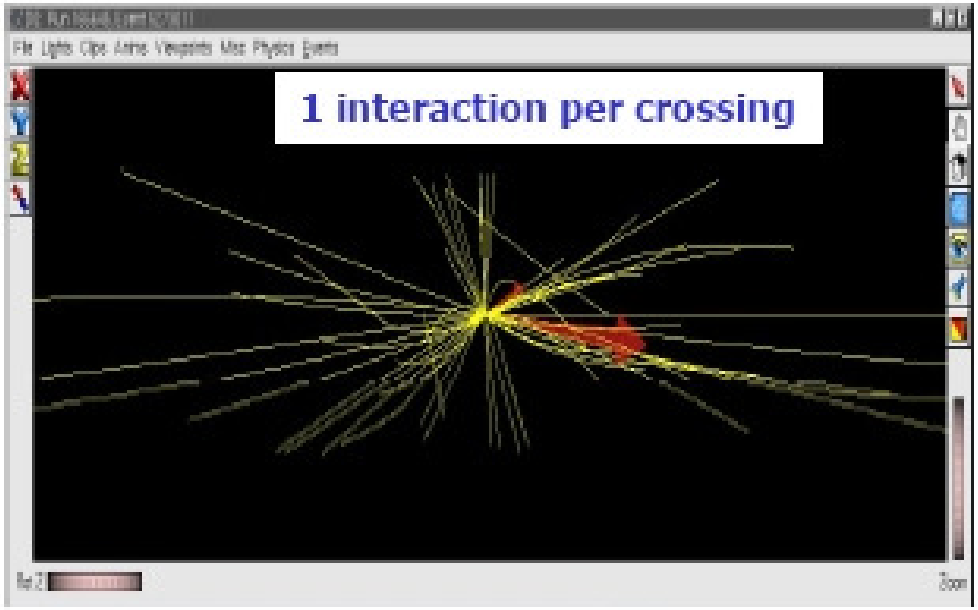
Level 1, 2 and 3 rates during typical Tevatron store



D0

Number of interactions per crossing



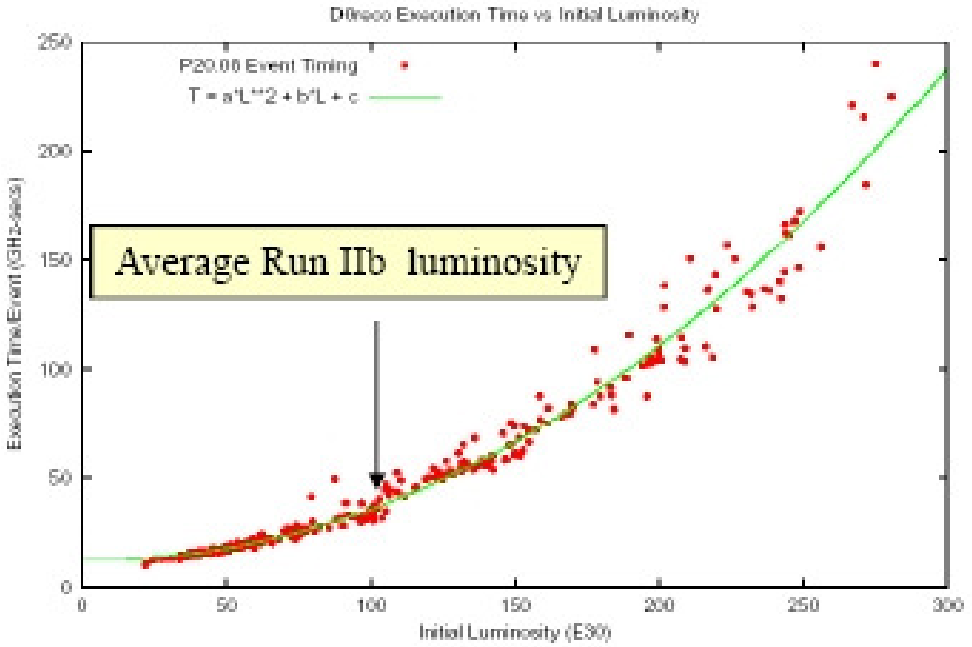


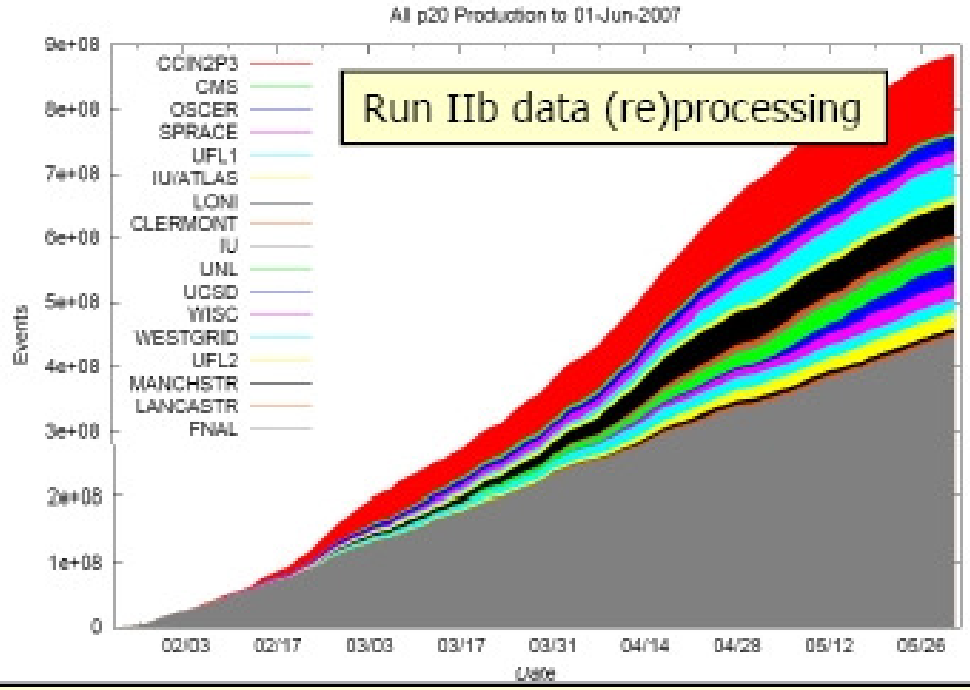
Z → ee events

Data Processing

- DØ collected over $2 \cdot 10^9$ events in Run II
- Current reconstruction program (Version p20) is in use since summer 2006
  - New Run IIb detectors
  - Preliminary Run IIb calibration data
  - Faster and more robust

For uniform and better quality data reprocessed Run IIb data collected before January 2007
Was accomplished on the GRID in ~4 months
Have full Run IIb data set available for analysis





Reconstruction timing

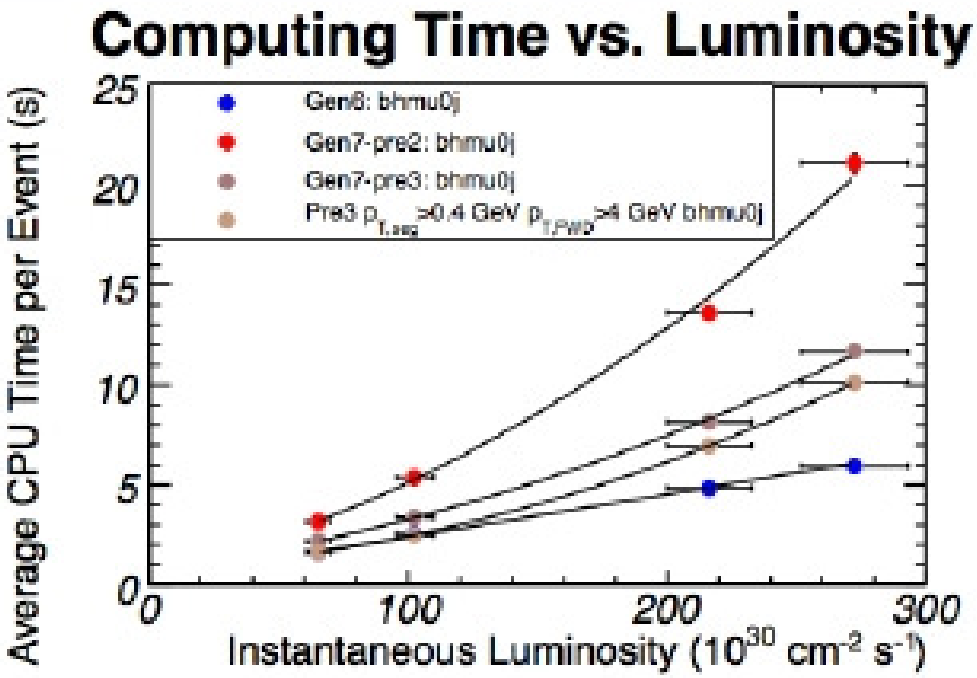
- Major time consuming process is tracking due to small number of tracking layers and high occupancies
- Currently reconstructing on the Fermilab's DØ farms ~5mln events per day and writing to tapes about the same number
- Within weeks extra computing resources will become available doubling available CPU
  - Will have "head room" and ready for even higher luminosity operation

DØ SAM based computing model and use of the GRID resources is a success!

Computing Time vs Luminosity

- Normalize luminosity curve to unit area in $[50\times10^{30}, 300\times10^{30}]$ cm^{-2} s^{-1}
- Parameterize CPU time curves with 2nd order polynomials
- Convolute the two curves: average CPU time per event

$$\langle t \rangle = \sum_{\text{lumi bins } i} t_i \cdot N_i \cdot \Delta\mathscr{L}_i$$



- Results:

Release	Average Time (s)	Ratio to Gen6
Gen6	2.72	1.00
Gen7-pre2	6.69	2.46
Gen7-pre3	4.12	1.52
Gen7-pre3, FWD p_T > 4 GeV/ c	3.42	1.26
Gen7-pre3, Segment p_T > 0.5 GeV/ c	3.33	1.23
Gen7-pre3, FWD p_T > 2, Seg p_T > 0.43	3.26	1.20

CDF & D0

1.5×10^{7} seconds/yr = 42 weeks/yr, 100 hours/wk of beam

Data taking efficiency 85%

FY	08	09	10
Total Int. Luminosity $f b^{-1}$	2.0	3.2	4.9	6.8	8.1?
Integrated Luminosity / yr	0.7	1.2	1.7	1.9	1.3?
pb^{-1} / wk delivered	16	28	50	50	50
pb^{-1} / wk recorded	14	24	42	42	42

Introduction: Some Numbers

	CDF	D0
Raw data size* (Kbytes/event)	150	250-300
Reco data size (Kbytes/event)	120	200
User format (Kbytes/event)	25-180	20-40
Reco time** (GHz-sec/event)	5(10)	50(120)
User analysis time (GHz-sec/event)	1(3)	1
Peak data rate (Hz)	130(360)	50(100)

*Raw event size depends upon trigger type and luminosity

**Reconstruction time depends upon raw data size

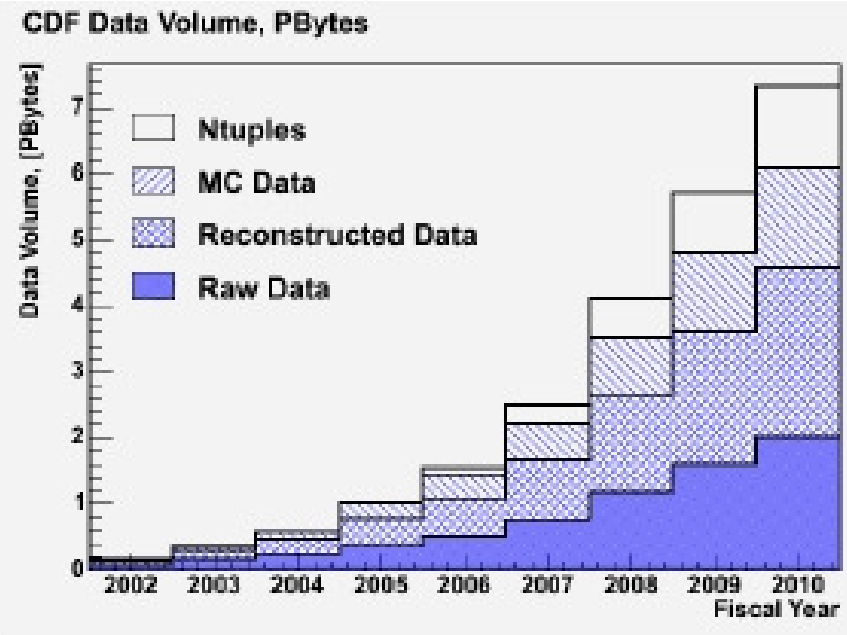
D0 Vital Statistics 2006

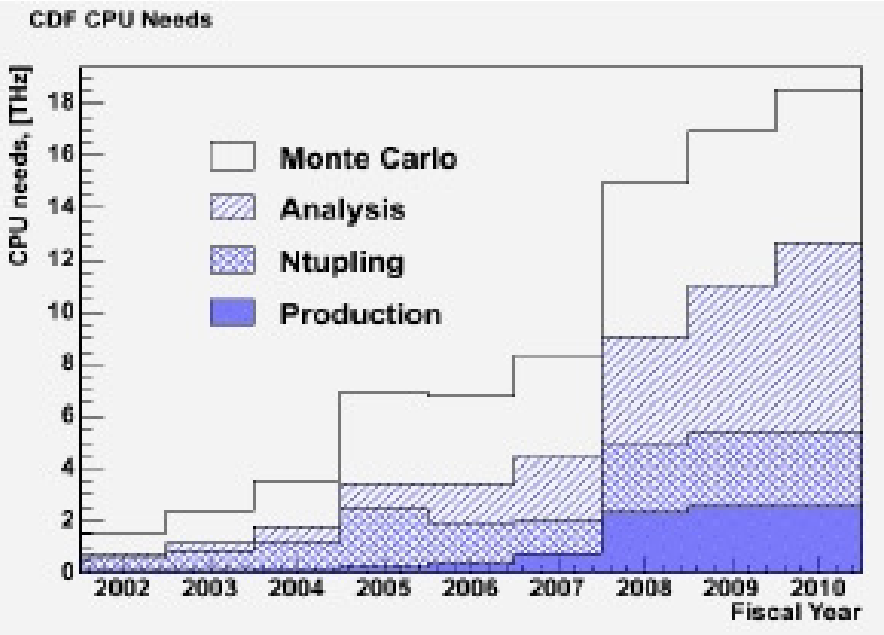
D0 Vital Statistics		
	1997(projections)	2006
Peak (Average) Data Rate(Hz)	50(20)	100(35)
Events Collected	600M/year	1.5 B
Raw Data Size (kbytes/event)	250	250
Reconstructed Data Size (kbytes/event)	100 (5)	80
User format (kbytes/event)	1	40
Tape storage	280 TB/year	1.6 pb on tape
Tape Reads/writes (weekly)		30TB/7TB
Analysis/cache disk	7TB/year	220 TB
Reconstruction Time (Ghz-sec/event)	2.00	50 (120)
Monte Carlo Chain	full Geant	full Geant
user analysis times (Ghz-sec/event)	?	1
user analysis weekly reads	?	3B events
Primary Reconstruction farm size (THz)	0.6	2.4 THz
Central Analysis farm size (GHz)	0.6	2.2 THz
Remote resources(GHz)	?	~ 2.5 THz(grid)

Challenge II: larger data samples

Computing model input parameters

Fiscal Year	2007	2008	2009	2010
Integrated luminosity (fb^{-1})	3.2	5.9	6.8	8.1
		4.9		
Total number of events (10^9)	5.0	8.0	11	13
Raw data logging rate (MB/s)	17	30	30	30





Computing inventory

1.5×10^7 seconds/yr = 42 weeks/yr, 100 hours/wk of beam

2006			Actual		Requirements		
		Fiscal Year	2007	2008	2008	2009	2010
	CPU (THz)	Estimated requirement			15	17	18
4.8		Fermilab	7.9	9.6	10	11	12
1.7		On-site contributions	1.7	1.7	1.7	1.7	1.7
2.3		Remote (dedicated)	1.6	1.6	1.6	2.3	2.3
		Opportunistic	1.7	1.7	1.7	2.0	2.0
8.8		Total available	13	15			
	Disk (PB)	Estimated requirement			1.0	1.3	1.5
0.6		Fermilab	0.7	1.0	0.98	1.2	1.4
0.2		On-site contributions	0.1	0.06	0.06	0.06	0.06
0.1?		Remote	0.1?	0.1?			
0.9?		Total available	0.9	1.2			
1.7	Volume on tape (PB)		2.6	—	4.1	5.7	7.3

CDF FY2007 procurements at Fermilab

- CPU
  - Shifted budget allocation from tapes into CPU
    - Tape cost dropped by 45% + lower than expected logging rate
  - Added net of 1.7 THz to CPU at Fermilab ($520k)
    (includes about $66k from Japan)
    - Will be available in November, 2007
      (much earlier delivery, deployment than past years)
- Disk
  - Replaced retirements in cache, expanded project disk, many new servers optimized for special uses ($350k)
- Tape drives
  - Added 7 LTO-3 drives for a total of 17 ($126k)

    Tape library cost of about $150k

PO		FY 04	05	06	07	Total
CDF	Disk	25	325	105	231	686
D0	Disk	24	195	330	264	813
						TB
CDF	CPU	950	960	1632	2480	15.7
D0	CPU	1200	880	1520	3280	17.9
				core	core x 2	T Hz

Starting FY07, each core is 2x speed of older cores, more Ops/Hz

core = 2.6 G Hz

> CDF Hardware systems:
Interactive Login Pool, groupCAF, FermiGrid,
and off-site CAF/Grid clusters

> Disk:

analysis	about	300 TB
diskpool	about	120 TB
dcache	about	350 TB
production	about	50 TB

> CDF disk (current capacity) ? about 800 TB

> distribution of the machines in terms of
> age, capacity, warranty status, hardware problems.

year	how much	warranty
FY-03:	105 TB	be decommissioned
FY-04:	80 TB	out of warranty
FY-05:	370 TB	under warranty
FY-06:	190 TB	under warranty
FY-07:	460 TB	under warranty

> How does CDF use their disk space ?

dcache auto-managed cache, in front of tape library
diskpool static, physics group control
fileservers physics group control for dataset assembly and static
project space under user control

> How much is tape backed dCache/sam cache ? about 350 TB

> How much is project space ?

about 500 TB < disk in FCC, not in the portcamps >

> What are the cache statistics ? not available
> dcache lifetime for files ? ~ weeks lifetime
> how often are files re-cached? not available

> What is the process wait time for pulling files from disk ?

depends on protocol, dccp read is slowest

> If the cache statistics are not available, what would it take to get them.

time, we are in the process of setting up age-gathering tools
re-cache statistics would need to come from dcache

> CDF CPU (current capacity) about 12,000 THz

> distribution of the machines in terms of age, capacity, warranty status.

year	number and type	warranty
FY-03:	242 dual ?	to be decommissioned
FY-04:	366 dual Intel Xeon 3.0 GHz	out of warranty
FY-05:	240 dual dual-core AMD Opteron 265, 1.8 GHz	under warranty
FY-06:	410 dual dual-core Intel Xeon 5148 2.33 GHz	under warranty
FY-07:	155 dual quad-core Intel Xeon 2.66 GHz	under warranty

> What is the pattern of usage for farm production, skimming,
root tuple production, significant analysis patterns.

> Networking:
no networking bottlenecks at this time,
we expect networking infrastructure to be final for CDF
(modulo a few 10Gb/s links and a switch upgrade or two)

> Tape: How much tape does CDF and D0 have
> and how is it distributed acrossthe robots.
don't know, Angela would have to dig this out

CDF

FY			06	07	08	09	10
Ave. Initial Luminosity		E30	120	220	280	280	280
Average Event Rate		Hz	100		200	200	200
Raw Data / event		KBytes	100	100	150	150	150
MBytes / sec	Data - to - Tape			17	30	30	30
CPU sec / event	Reconstruction		2	3	5	5	5
CPU sec / event	Ntupling			6	10	10	10
Recon Data / event		KBytes	120				
User Data / event		KBytes	180				
Total number of events		10^9		5	8	11	13
CPU		THz	4.8	7.9	10	11	12
Disk		Pbytes	0.6	0.7	1.0	1.3	1.5
Tape		Pbytes	1.5	2.6	4.1	5.7	7.3

D0

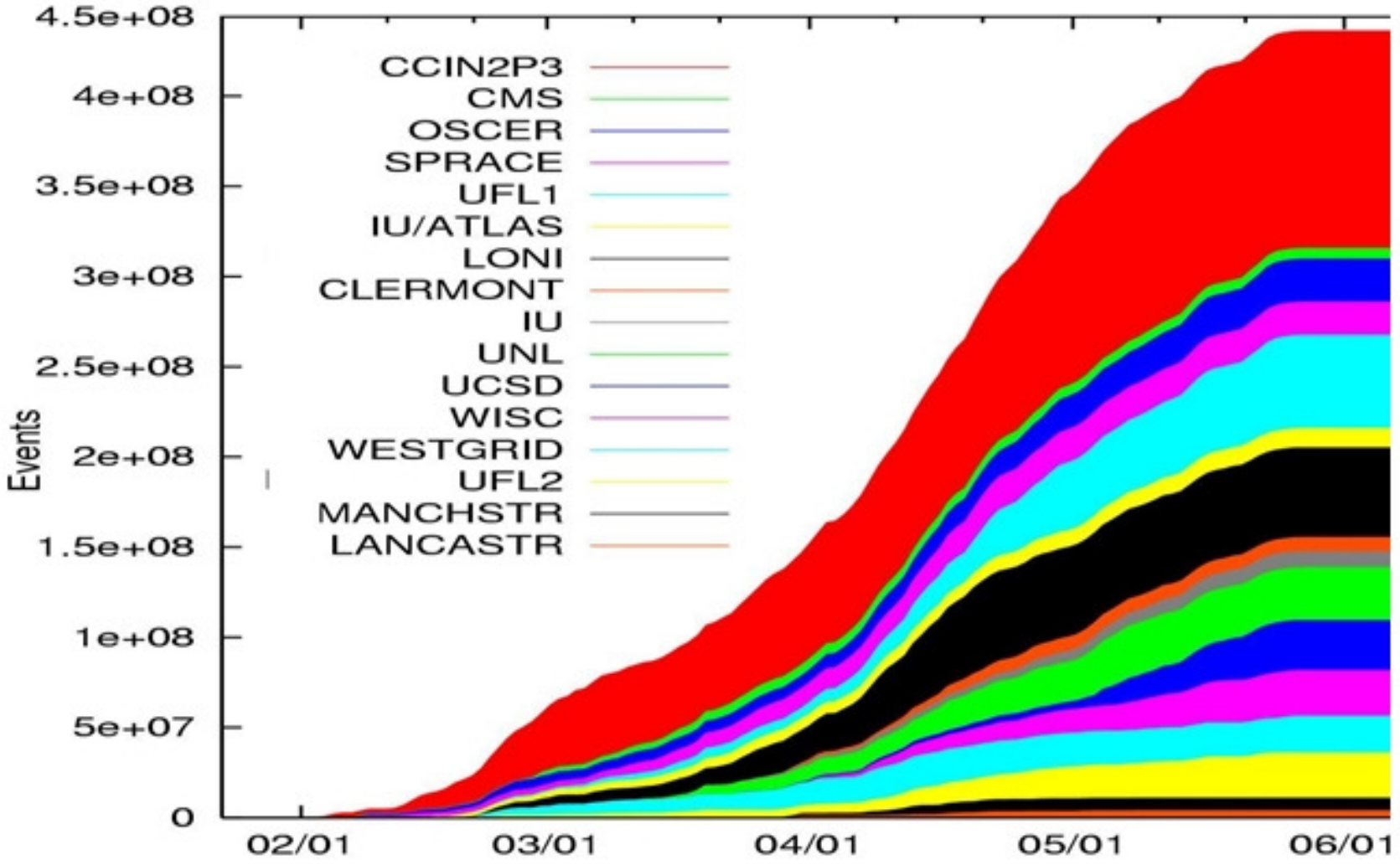
FY		06	07	08	09	10
Ave. Initial Luminosity	E30	120	220	280	280	280
Average Event Rate	Hz	35		100	100	100
Raw Data / event	KBytes	250	250	300	300	300
MBytes / sec Data - to - Tape		9		30	30	30
CPU sec / event Reconstruction		20		46	46	46
CPU GHz sec / event Recon		50		120	120	120
Recon Data / event	KBytes	80	100	100	100	100
User Data / event	KBytes	40	50	50	50	50
Total number of events	10^9	2	3	4.5	6	7.5
CPU	THz	4.6				12
Disk	Pbytes	0.55	0.8			1.5
Tape	Pbytes	2	2.x	3.5		7.3

D0 2007 Data Re-Processing on OSG

- Improved detector understanding and new algorithms require re-reprocessing of the raw detector data
- Input: 90Tb of detector data + 250 Tb in executables
- Output: 60 Tb of data in 500 CPU years
  - DZero did not have enough dedicated resources to complete the task in the target 3 months

- D0 requested OSG to provide 2000 CPU for 4 month.

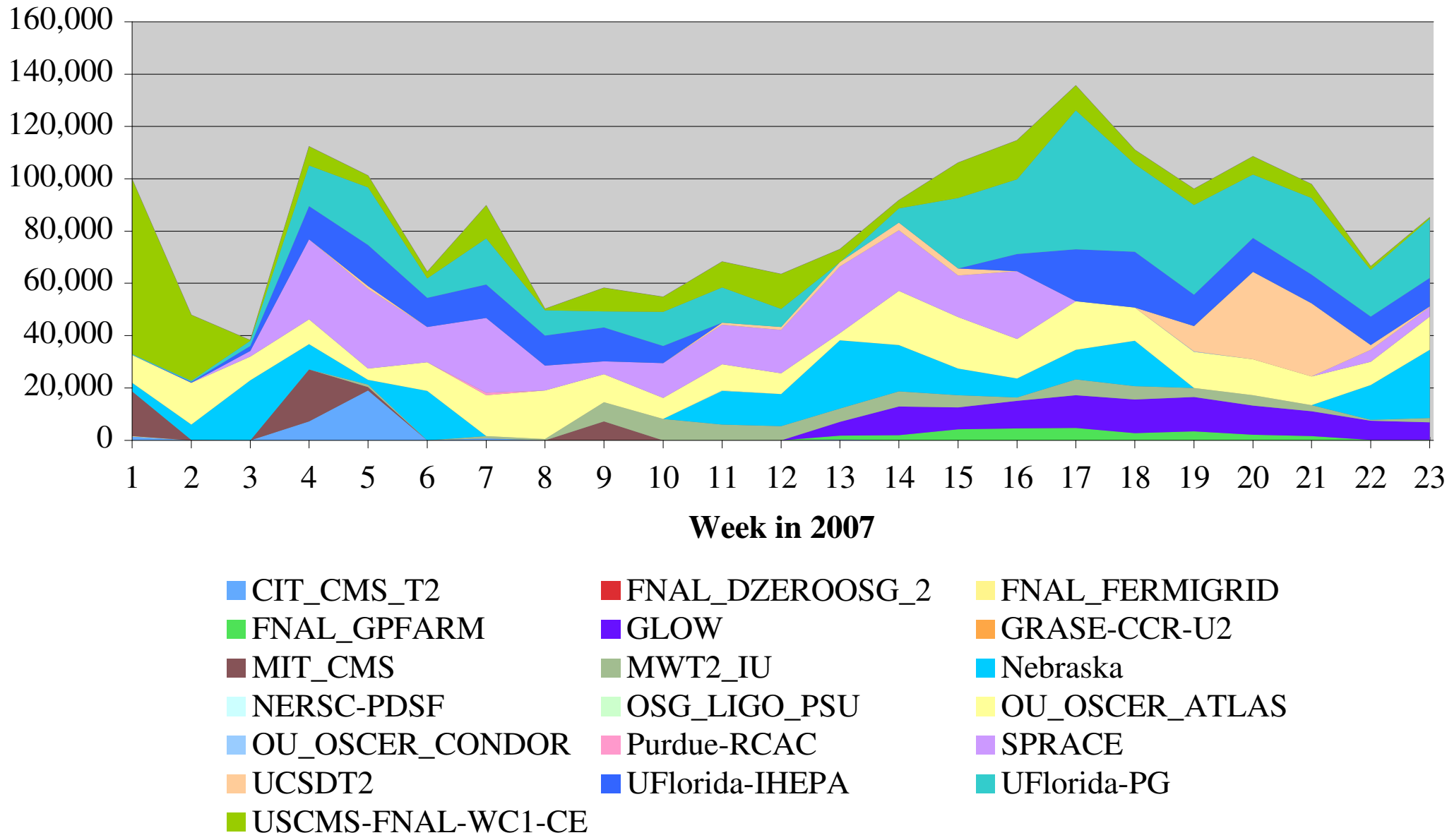
D0 2007 Data Re-Processing on OSG



2007

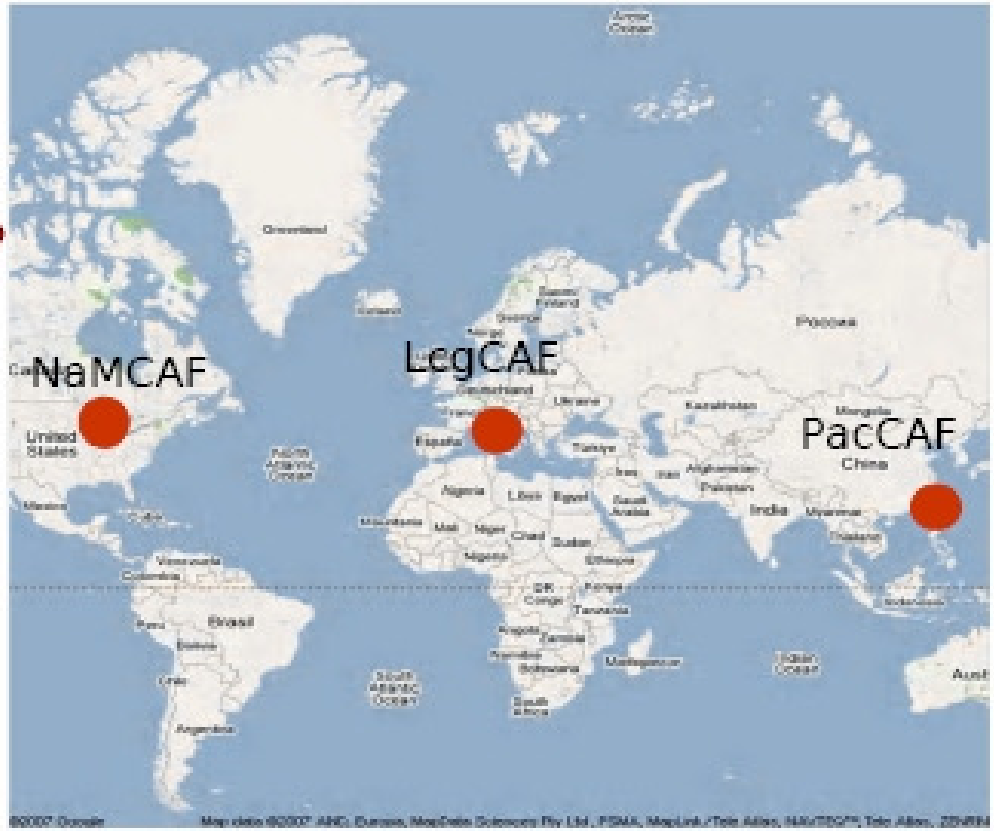
- 450M collider events delivered to physicists
  - Reconstructed in fully distributed, opportunistic environment

D0 OSG CPU_hour/week



CDF

Towards GRID



Eliminate all but three grid submission portals: NAMCAF, LcgCAF, PacCAF.

Migrate all existing systems accordingly. (May keep FermigridCAF+CNAF for data access.)

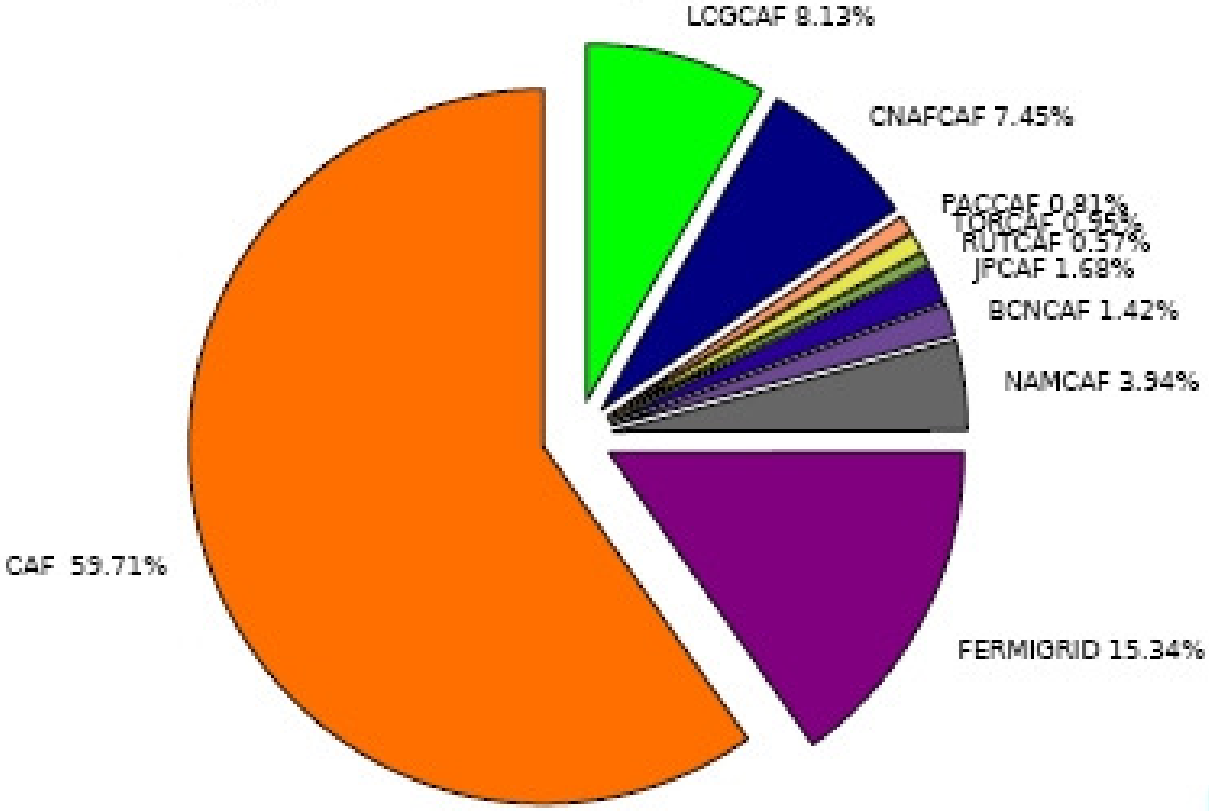
Current CDF Dedicated Resources

Current Resources [*]			
Cluster Name and Home Page	Monitoring and Direct Information Links	CPU (GHz)	Disk space (TBytes)
Original FNAL CAF	queues, user history, analyze, ganglia, sam station, consumption	1000	370
FNAL CondorCAF (Fermilab)	queues, user history, analyze, ganglia, sam station, consumption	2200	(shared w/CAF)
CNAFCAF (Bologna, Italy)	queues, user history, analyze, resources, network, sam station, datasets, consumption	480	32
KORCAF (KNU, Korea)	queues, user history, ganglia, sam station, datasets, consumption	178	5.1
ASCAF (Academia Sinica, Taiwan)	queues, user history, ganglia, sam station, datasets, consumption	134	3.0
SDSC CondorCAF (San Diego)	queues, user history, analyze, ganglia, sam station, datasets, consumption	380	4.0
HEXCAF (Rutgers)	queues, cpu, sam station, datasets, consumption	100	4.0
TORCAF (Toronto CDF)	queues, user history, analyze, ganglia, disk status, sam station, datasets, consumption	576	10
JPCAF (Tsukuba, Japan)	queues, user history, ganglia, sam station, datasets, consumption	152	10
CANCAF (Cantabria, Spain)	queues, user history, ganglia, sam station	50	1.5
MIT (Boston, USA) (MC only)	queues, user history, analyze	322	3.2
	Current Totals (*):	5572	448

http://www-cdf.fnal.gov/internal/fastnavigator/fastnavigator.html (2006/Aug)

Resources Usage: All Farms

Averaged over a year: Jan07-Oct07

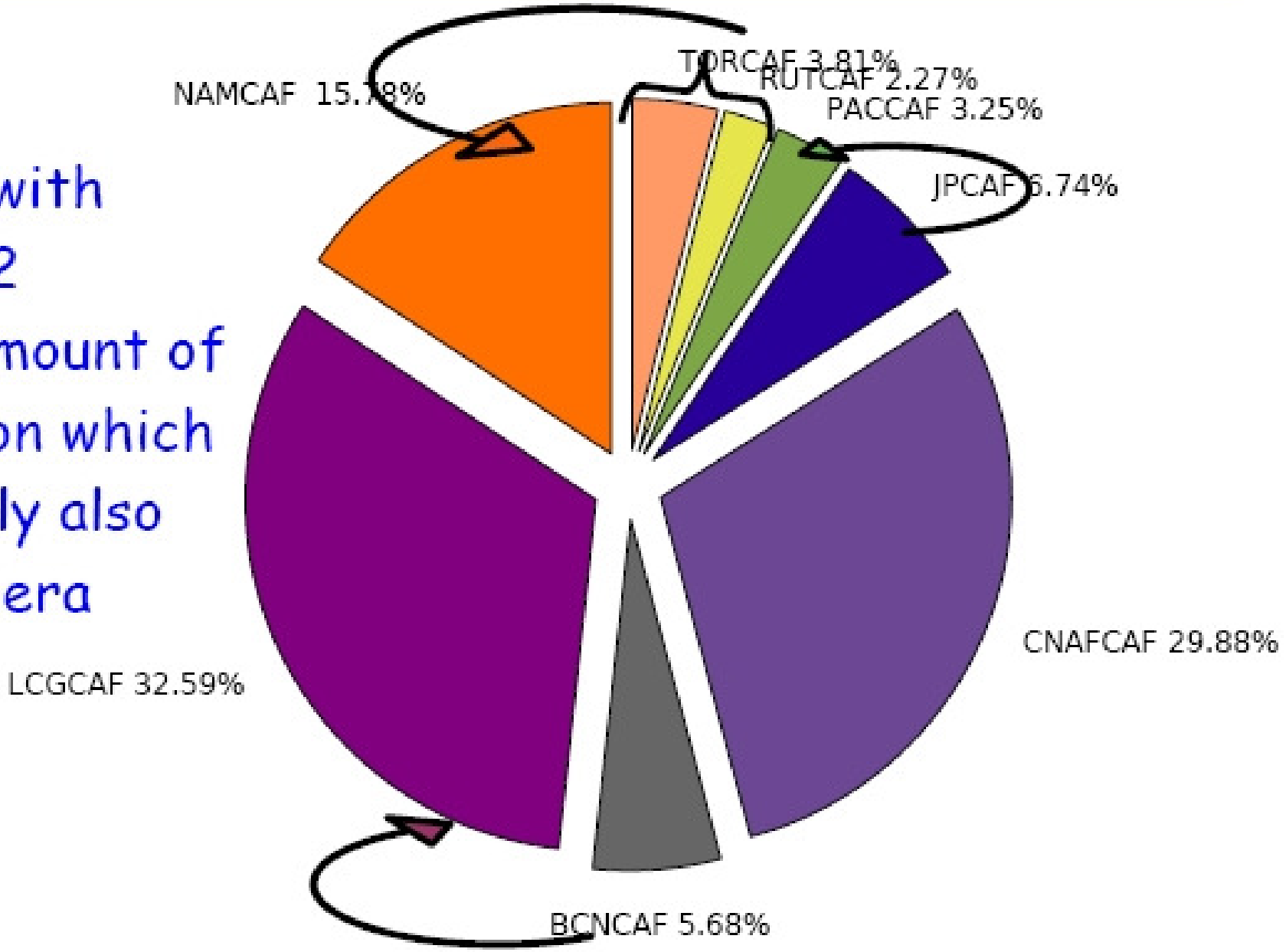


This is an "evolving pie": thanks to FermiGrid use Aug-Oct onsite share has been 50%:50%

Quite soon:
- dcafs will disappear
- caf will be merged in FermiGrid

Offsite Resources Development

Negotiate with Tier1/Tier2 a certain amount of resources on which CDF can rely also in the LHC era



On-site Resources Use

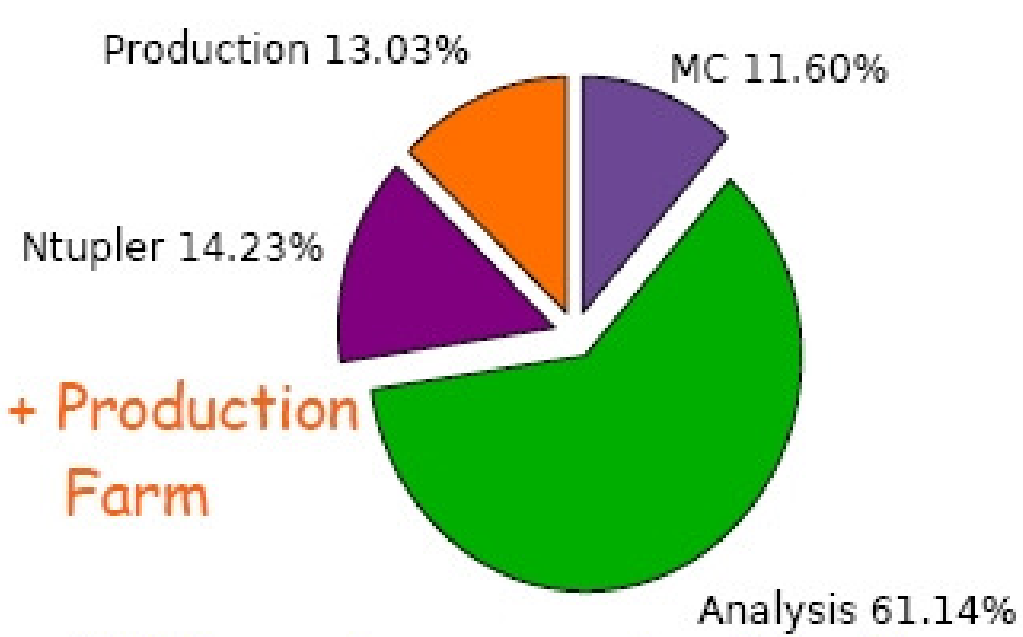
Production: Productionexe

MC: "CDFSim"

Ntupler: STN+top+Bs ntuple

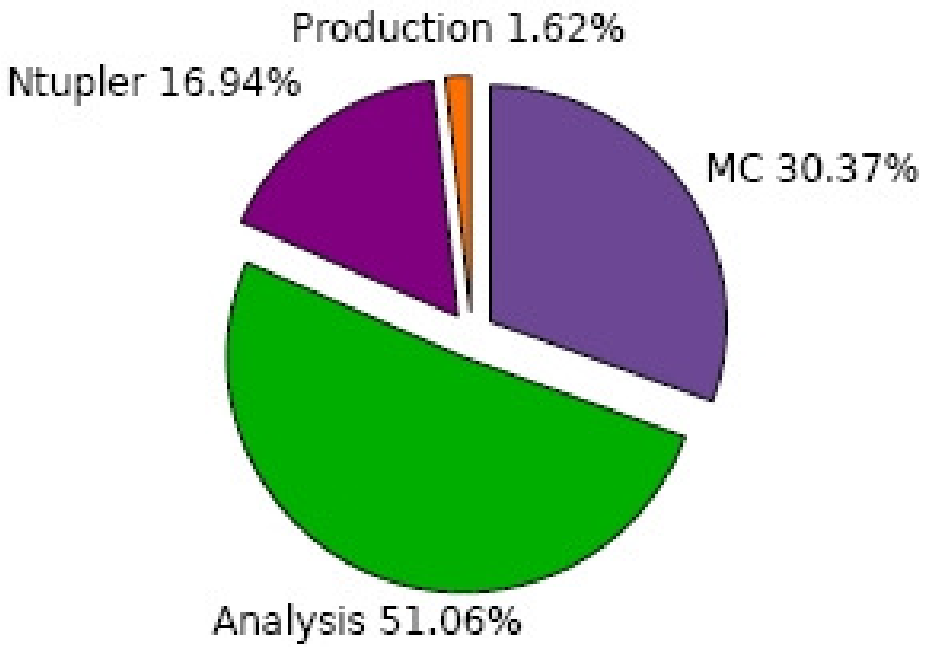
Analysis: all the rest

CAF



+ Production Farm

FermiGrid



MC has to moved off-site

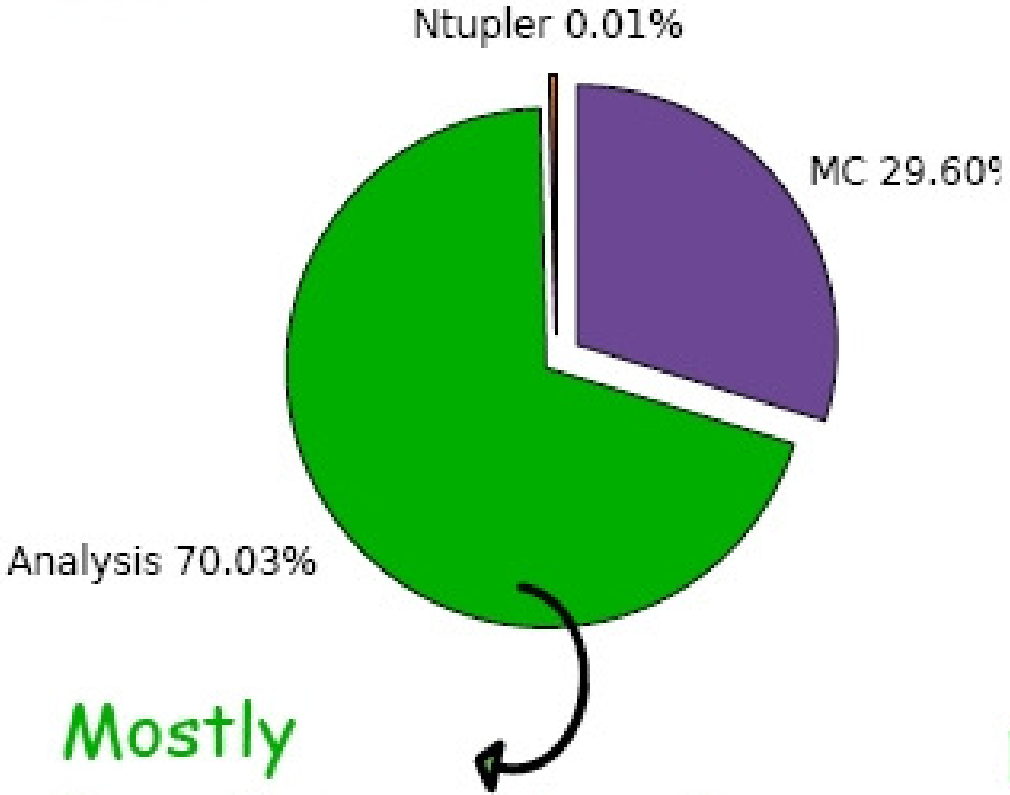
On-site resources mainly dedicated to data production & analysis

Off-site Resources Use

Production: Productionexe
MC: "CDFSim"

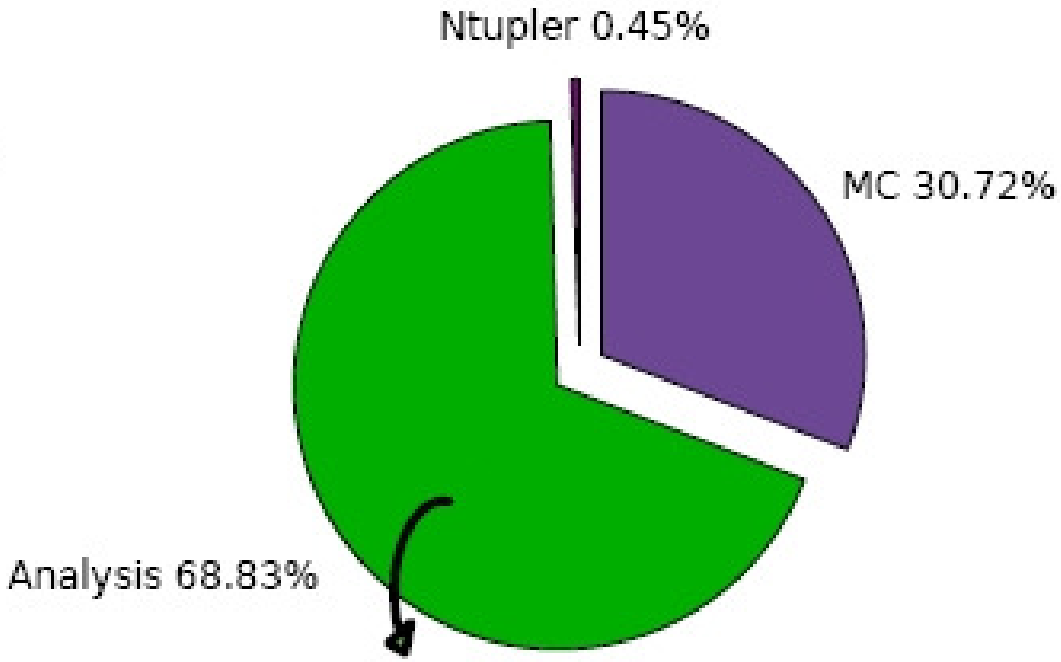
Ntupler: STN+top+Bs ntuple
Analysis: all the rest

dCAF



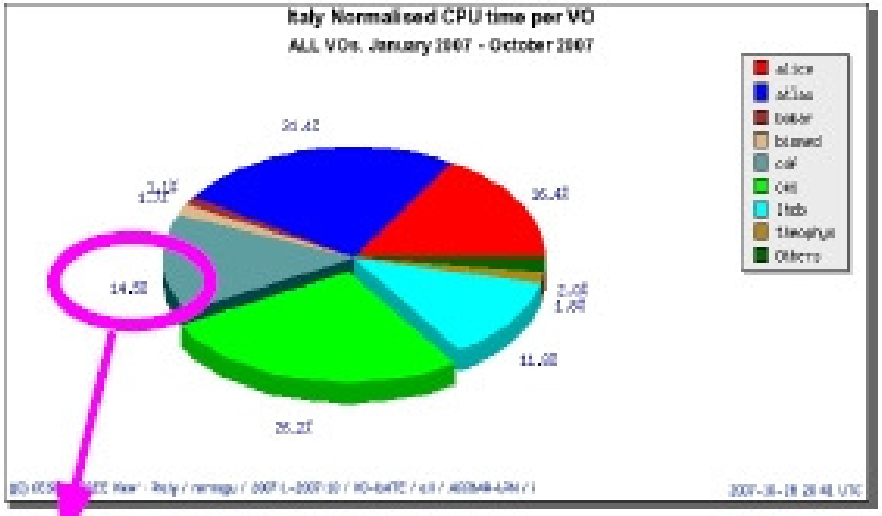
Mostly
Pseudo-experiments

CNAFCAF

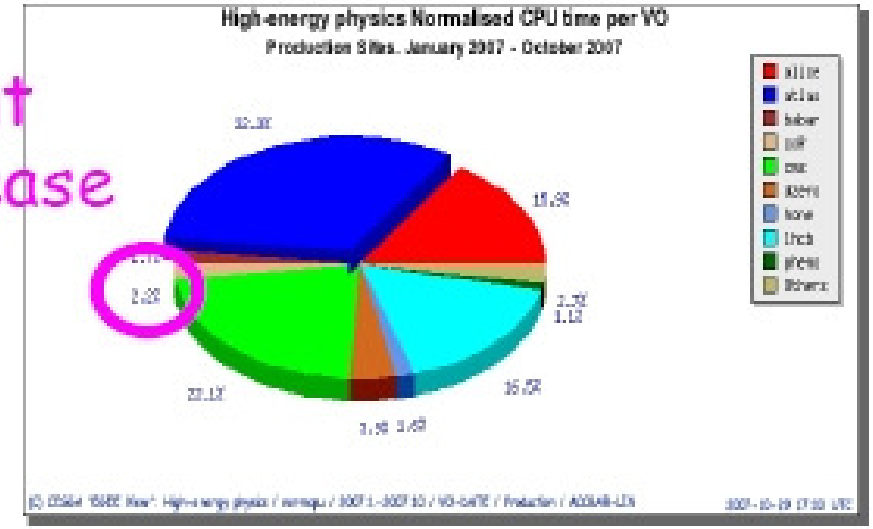


Not negligible contribution
of user analysis

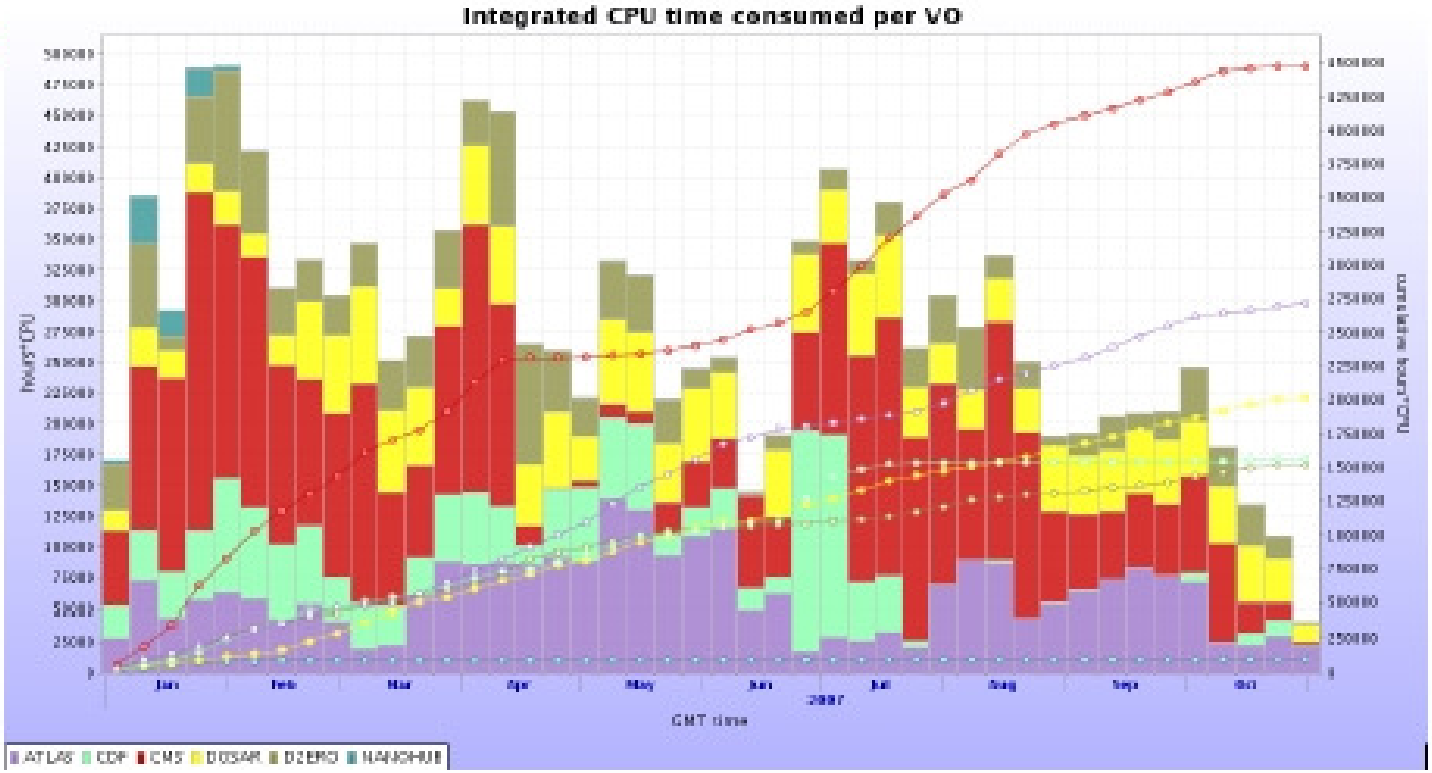
VO Use



We want to increase



As big as one LHC experiment



Future Challenges

- Higher instantaneous luminosity
  - Larger events, slower reconstruction, tracking more difficult, need more CPU per event
- Higher integrated luminosity and higher data taking rate
  - Larger data samples
    - Need more processing power
    - Need more storage
- Migration of physicists to LHC experiments
  - Human resources for operations are shrinking
- FY2010 Running has been proposed

Additional Info/slides

	Effort Report "FTE"		Resource Needs "FTE"	
	2005	2006	2007	2009
Operations	107	91	68	68
Computing	35	30	32	25
Algorithms	74	65	55	21
Management	14	14	10	10
Total service contributions	230	200	165	124

Conclusions

- The DØ detector is working well with high data taking efficiency
  - Currently 2.5 fb^{-1} on tapes
  - No major technical issues to continue data collection up to and above 8 fb^{-1}
- Data processing is keeping pace with data collection, MC production is steady

CDF Resources available

	CY 2007	2008	2009
US FTE	222	162	127
Non US FTE	170	135	109
Total US + NonUS	**392**	**297**	**236**
Post Doc's	101	73	53
Students	147	102	77

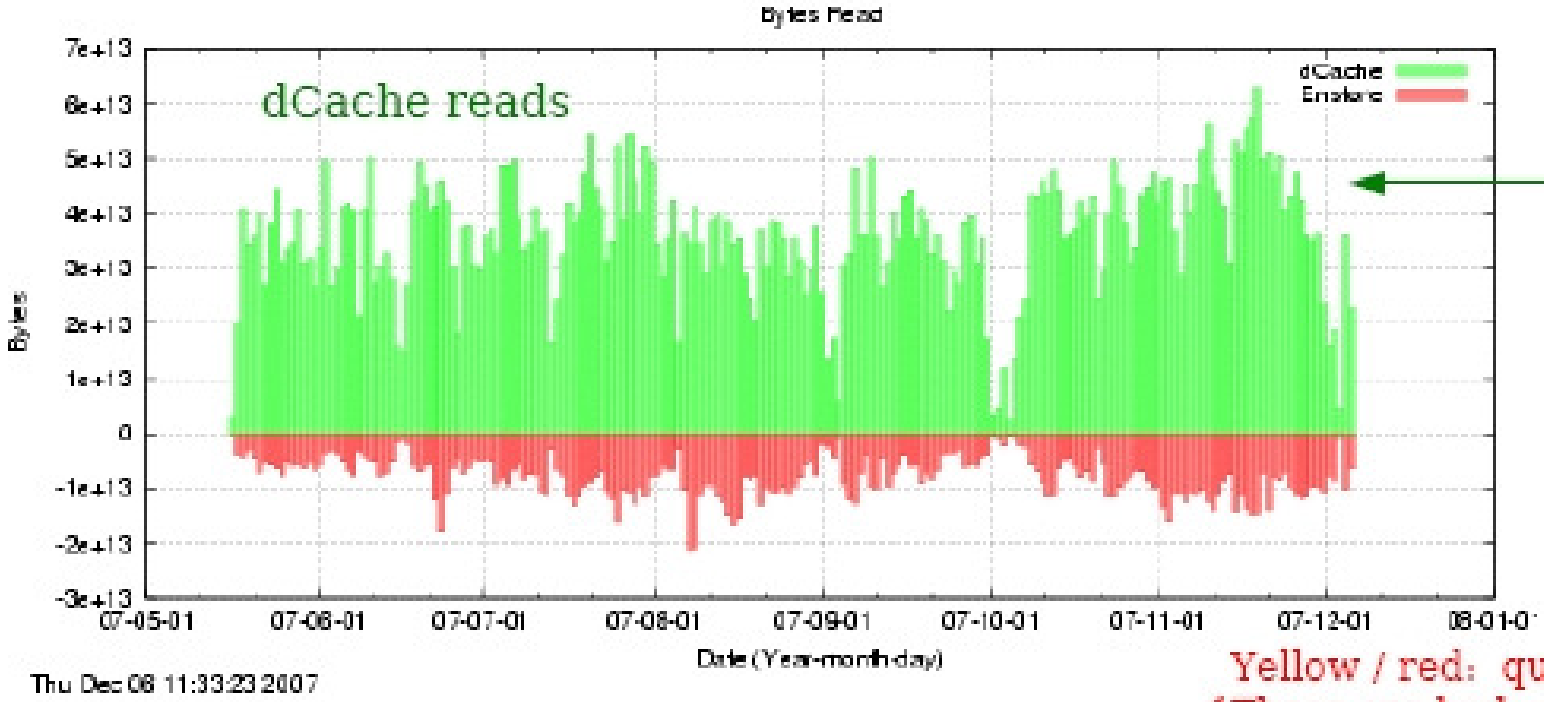
Collaboration members available in units of FTE

~25% more FTE in CY07 than estimated in 2005

It takes ~100 FTE to Run CDF

Operations

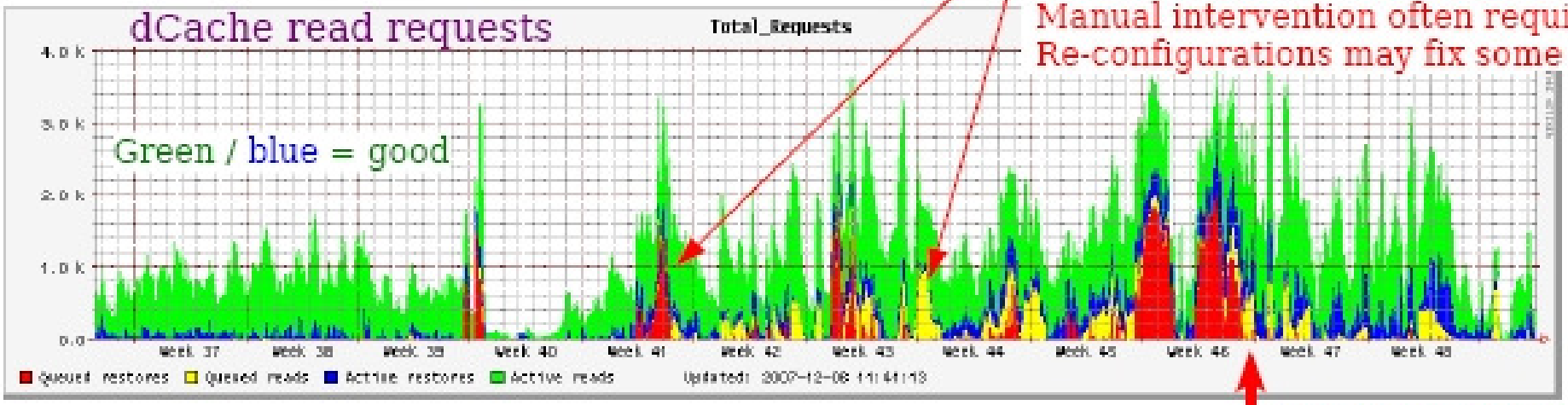
- **Production processing** (R. Culbertson, E. Gerchtein, R. Harr, B. Jayatilaka, T. Miao, M. Vogel, A. Warburton + calibrators, ntuplers, MC producers)
  - Raw data / ntuple production proceeding on schedule
    - P13 raw data production completed, ntuples almost done
    - Infrastructure / error handling improvements over past few months
      - Processed P13 at record rates (>40 M events/day)
      - Concatenation throughput higher than in the past.
      - Working to further reduce the time for recoveries, clean-up
    - Start P14 after calibration sign-off in about 2 weeks
  - MC production
    - Problem with latest tarball (patch J) delaying P13 MC
      - Expect a resolution within days



Increased read rate over past month

Yellow / red: queued read / tape restores
These are bad

Manual intervention often required
Re-configurations may fix some cases



Fixed tape queuing(?)

Installed Enstore Systems

Enstore provides distributed access to and management of data stored on tape. It provides a generic interface so experimenters can efficiently use mass storage systems as easily as if they were native file systems.

STKEN Enstore System	Mass Storage Production Service for General Fermilab Users
CDFEN Enstore System	Mass Storage Production Service for CDF Run II
DOEN Enstore System	Mass Storage Production Service for D0 Run II
GCCEN Enstore System	Mass Storage Internal Testing/Debugging
Production System's Overall Status	Status for all Production Enstore systems

Total User Data on Tape (Cdfen, D0en, Stken) :	7996.090 TB

http://rexganglia1.fnal.gov/

Available ganglia pages:

Farms (CDF, D0, and GP)

CDF Offline

CDF Online (requires login)

D0 Offline

D0 Online

MINOS

Cluster Report for Wed, 6 Feb 2008 05:43:46 -0600

Get Fresh Data

Metric: load_one Last: hour Sorted: descending

Physical View

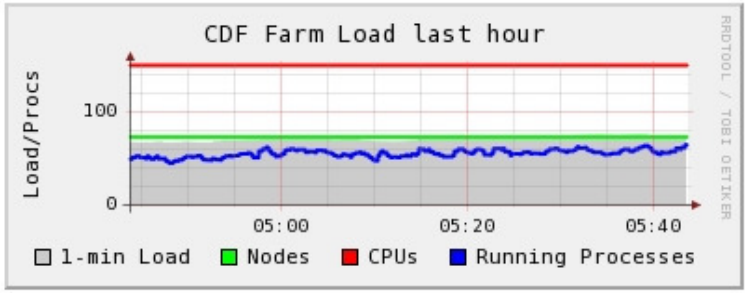
Farms Grid > CDF Farm > --Choose a Node

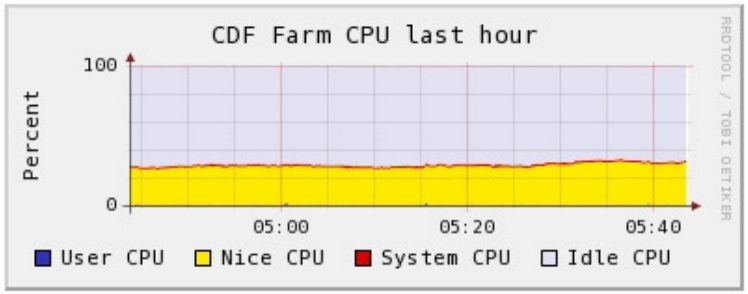
Overview of CDF Farm

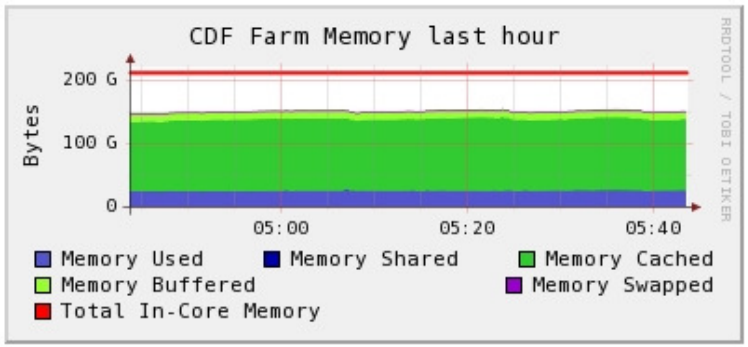
CPUs Total:	**150**
Hosts up:	**73**
Hosts down:	**3**

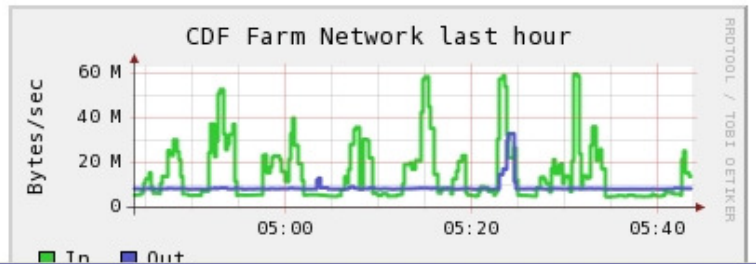
Avg Load (15, 5, 1m): **46%, 49%, 48%**

Localtime: **2008-02-06 05:43**









Pie Chart

Fermi National Accelerator Laboratory

FEF Faultlog

D0	CDF	FEF	GP Grid	MiniBoone
MINOS	MIPP	SCIBOONE		

Overview

- Ganglia thinks 16 nodes are down. (3275 up) View all.
- Of those 16 machines, 0 outages have been acknowledged*. View all.
- 908 nodes are not reporting to Ganglia.
- 4199 total nodes found in SYSADMIN database.

Outages

- 783 entries in faultlog with a recorded outage duration.
- Of those, the shortest outage was for the host FND0749, with a duration of 3 minutes.
- The longest outage was for D0OL95, and lasted for 153 days.
- Average outage is 9 days.

Queries

- See activity for past seven days.
- Generate a list of machines not reporting to syslogDB (slow link, takes ~2 minutes to load)
- Locate a single node: (partial names accepted)

Search

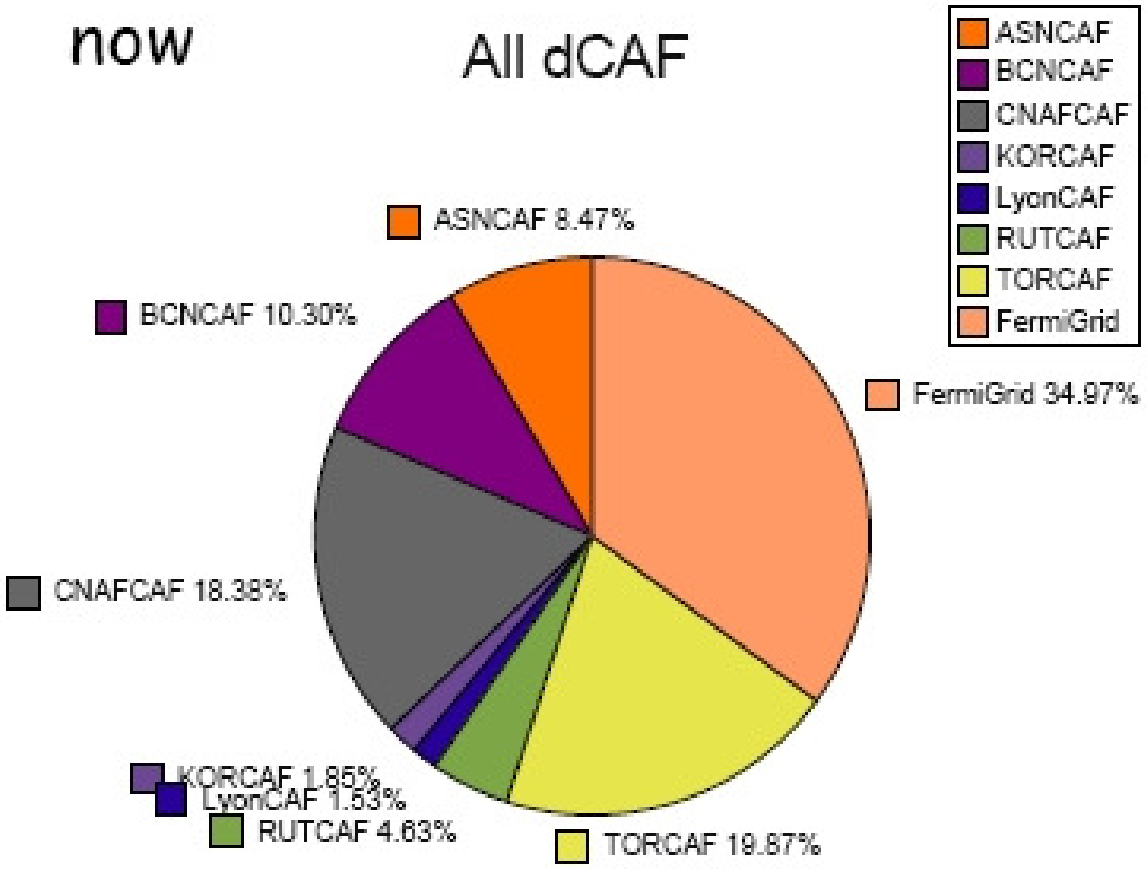
Update Frequency

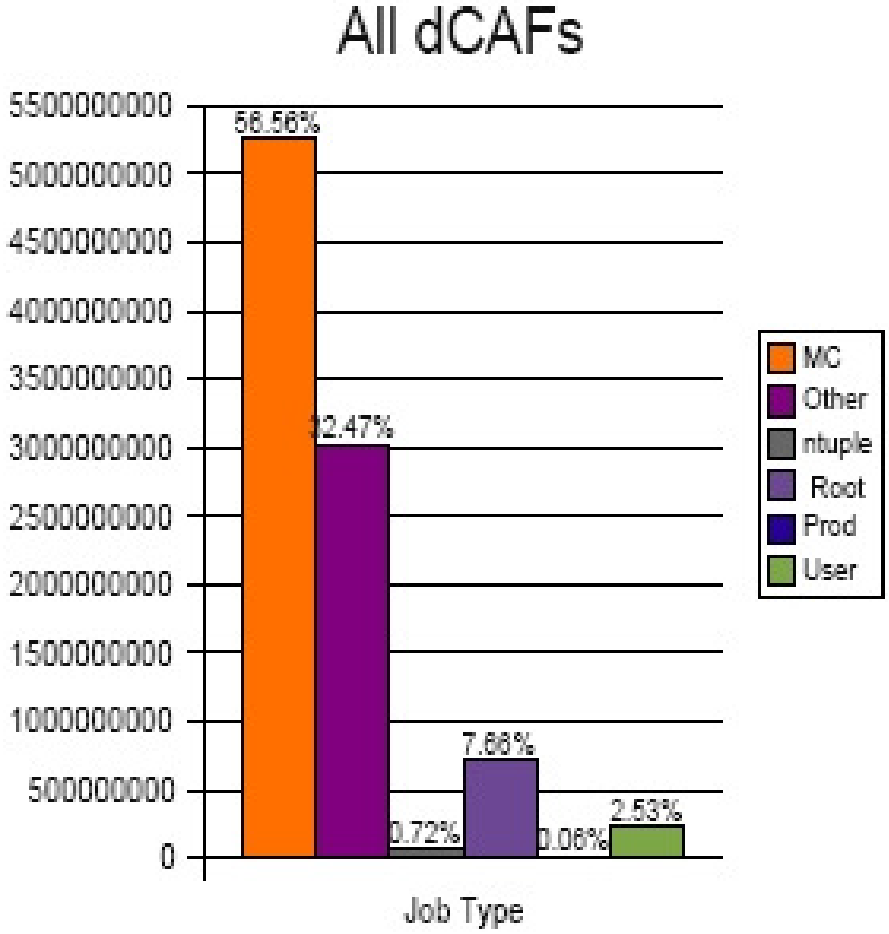
- Ganglia data updates once per minute.
- Hardware calls update once per hour.
- Cluster information updates once per day, early in the morning.

* Acknowledged is defined as a faultlog entry that is more recent than the last time the machine reported to ganglia.

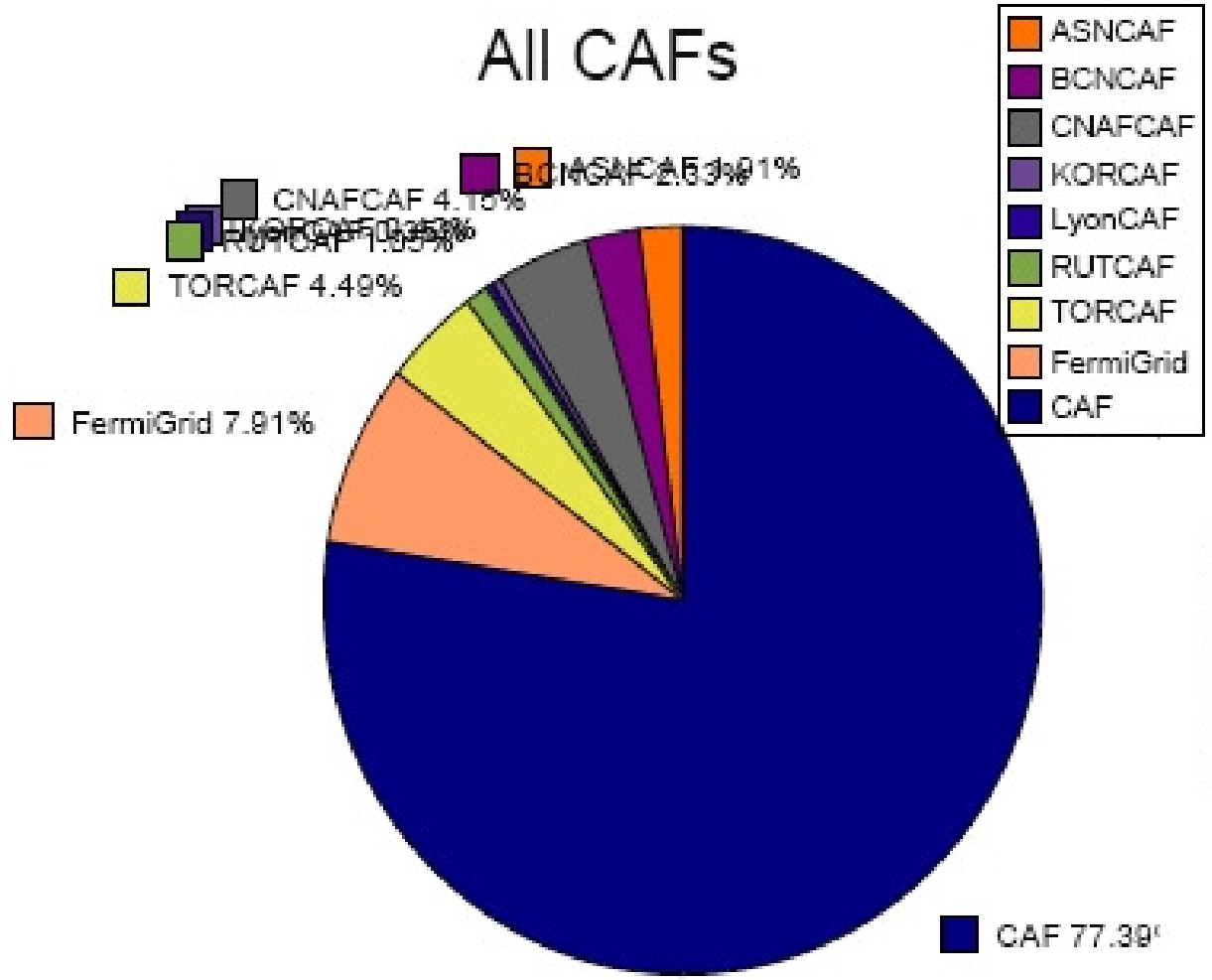
Usage of Dedicated Farms: all dCAFs

rtime from September 06 up to now



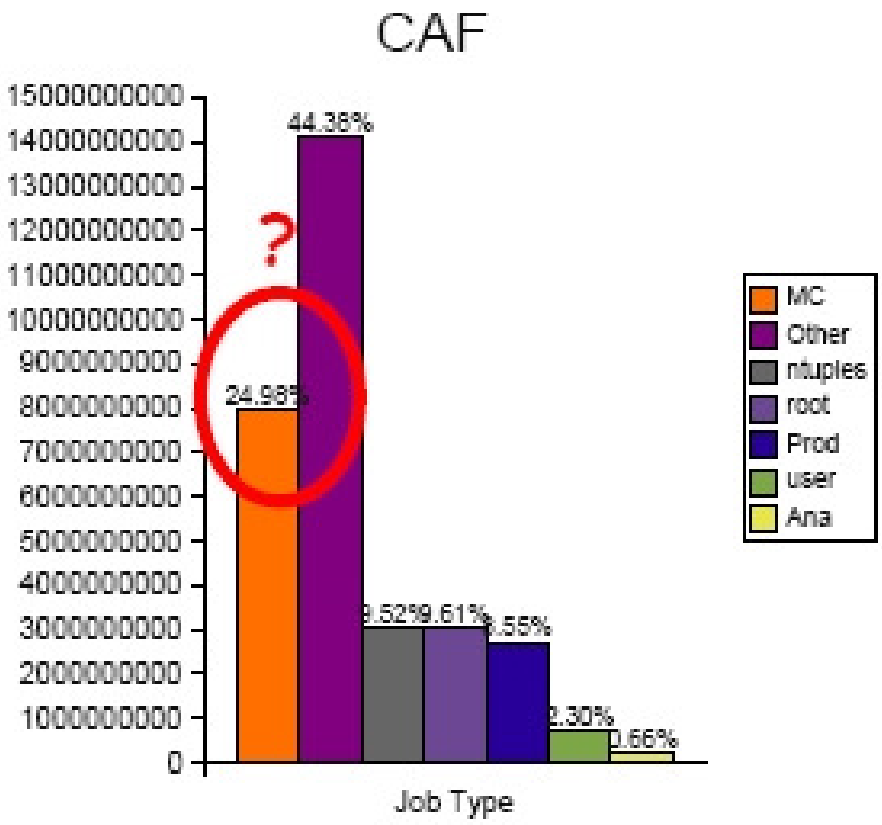


Usage of Dedicated Farms: CAF



CAF is a great system but too overloaded!

Policy: "Use CAF only for data"



Being overloaded contributed to the recent failures

Thanks

Amber's Starting List

Amber's presentation
www-hep.uta.edu/~d0race/d0rac-wg/amber-future-budget.ppt

Andy gave me Lucchesi/Snider "Offline status and plans"
12/2007 presentation to CDF Executive Board
http://www-cdf.fnal.gov/internal/WebTalks/
http://hcp2006.phy.duke.edu/HCP2006-science.html
D0 Computing and Analysis Model by Amber
CDF Computing and Analysis Model by Pierre Savard

Jason Allen gave me 2 useful D0/CDF links:
http://rexganglia1.fnal.gov
http://d0om.fnal.gov/d0admin/faultlog/

Roman Lysak
enstore (tape usage at CDF/D0) + dCache
http://www-ccf.fnal.gov/enstore/
network:
fndcg0.fnal.gov/~netadmin/nwm/cgi-bin/temp/core.html
local CDF CAF and CDF farms outside Fermilab:
http://cdfcaf.fnal.gov/

Paris CDF Week Collaboration Meeting
http://lpnhe-cdf.in2p3.fr/cdf_parismeet/

Roser	
Glenzinski	Preparing for summer conferences
Hahn	Detector Operations Status
Lucchesi	Offline Operations Status
Moore	Accelerator Status and Plans
Nurse	Trigger and High Luminosity

http://cdorg.fnal.gov/rex/status%20report/20070416/20070416.htm

Initial luminosity & Integrated Luminosity
http://www-cdfonline.fnal.gov/ops/opshelp/stores/

www-cdf.fnal.gov/internal/WebTalks/Archive/0712/071205_joint_physics/

CDF Computing Highlishts, Status, and Plans
www-cdf.fnal.gov/physics/ifc/2007-10-30/donatella.pdf

CDF Computing model and budget
www-cdf.fnal.gov/physics/ifc/2007-10-30/snider.pdf

CDF Monte Carlo Production on LCG via LcgCAF Dec. 2007
www.pd.infn.it/~lucchesi/talks/escience-lcgcaf.pdf

CDF Offline status and plans
www-cdf.fnal.gov/internal/WebTalks/Archive/0712/071206_cdf_exec_board/